Estimating causal effects with non-experimental data

by

David Lenis

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

July, 2017

Abstract

In this manuscript we seek to relax some of the traditional assumptions associated with the estimation of causal effects. In particular, we relax the assumption that all confounders are measured without error and the assumption that the observations in the sample are independent and identically distributed. Furthermore, we explore the impact of model misspecification in the estimation of population causal effects.

Advisor: Elizabeth A. Stuart

Committee: Kitty Chan, Bryan Lau, Carol Thompson

Alternates: Tom Louis, Karin Tobin

Acknowledgments

To my advisor Elizabeth Stuart, thank you for your mentorship and generosity during the past five years. Your support has been invaluable.

To my collaborators Benjamin Ackerman, Nianbo Dong, Cyrus Ebnesajjad and Trang Nguyen, thank you for the opportunity to address interesting questions together.

To Karen Bandeen-Roche and Marie Dinner-West, thank you for your support and the opportunity to teach in Armenia.

To Jimmie Lou DeBakey and my Sommer Scholar family, thank you for your love and support.

To the members of the Measurement Error working group, Ryan Andrews, Frauke Kreuter, Hwanhee Hong, Vivek Khatri, Kara Rudolph, Ian Schmid and Yenny Webb Vargas, our discussions enriched my work.

To my classmates and friends, I have learned so much from you. In particular I want to thank Francis Abreu, Vivek Charu, Caroline Epstein, Jean-Phillipe Fortin, Emily Huang, John Muschelli, Helen Powell, James Pringle, Tianchen Qian, and

ACKNOWLEDGMENTS

Elizabeth Sweeney,

To my friends from the School of Public Health, Tyler Alvare, Valery Caldas, Lillian Collins, Sally Dunst, Laina Gagliardi, Simi Grewal, Nicholas Khan, Hilary Samples and Jennifer Sherwood their commitment to make the world a better place inspires me daily.

To Mary Joy Argo, Mark Chiveral, Deborah Cooper, Anna Dent, Marti Gilbert, Patty Hubbard, Ashley Johnson and Debra Moffitt. I could have not done this without your invaluable help and kindness.

To my family, thank you for your unconditional support.

To my father.

Contents

CONTENTS

List of Tables

List of Figures

LIST OF FIGURES

Chapter 1

Introduction

At its core, the goal of causal inference is to identify causal relationships and distinguish them from a correlation analysis.

Estimation of causal effects is of vital importance in public health. Accurate estimation of causal effects allows researchers and practitioners to assess public policy, compare treatment efficacy, and evaluate health disparities. In this context, the main contribution of Donald Rubin and his collaborators was the systematization and formalization of key concepts that evolved in a concise framework, based on which causal effects could be defined. This conceptual framework is referred as the Rubin Causal Model (RCM).

Based on the idea of potential outcomes (Neyman, 1923), the RCM outlined and listed a precise set of assumptions in which causal effects can be estimated. Before presenting some key aspects of the RCM, some notation needs to be introduced.

Throughout this manuscript, the outcome of interest is represented by the letter Y, the letter T represents a binary exposure/treatment, the set of confounders (i.e. a set of covariates that are associated with the outcome and the exposure) is identified by $\mathbf{X}$. For each unit i, $Y_i(t)$ with $t = 0, 1$ represents the outcome that would have been observed if unit i received the treatment t. For any unit i, the individual level causal effect is defined as $Y_i(1) - Y_i(0)$. Notice that only one potential outcome in the pair $\{Y_i(0), Y_i(1)\}$ is observed, therefore individual causal effects cannot be estimated. Thus, a common practice is to estimate average causal effects.

Traditionally, the two average causal effects most commonly estimated are: (1) the average treatment effect (ATE) and (2) the average treatment effect on the treated (ATT). The ATE is defined as the average of the individual treatment effects over the population, whereas the ATT is defined as the population average of the individual treatment effect among those units who were actually treated. One of the main contributions of the RCM is the identification of the required assumptions to correctly identify and estimate the ATT and ATE. The key assumptions needed to estimate causal effects can be summarized as follows:

1. **Consistency** of the observed outcome. As previously stated, for any given unit i the pair $\{Y_i(0), Y_i(1)\}$ for $i = 1, 2, ...n$ is not observable, we only observe one of the two potential outcomes. In other words, the observed outcome Y_i is equal to $Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i)$. Thus for any unit, the observed outcome is equal to the its potential outcome under the treatment received.

2. **Ignorability** assumes that $\mathbf{X}$ contains all possible confounders. In other words, given the set of observed covariates $\mathbf{X}$, the treatment assignment is independent of the potential outcomes. The ignorability assumption implies that the treatment assignment can be assumed to be random, conditionally on observed characteristics of the units in the sample. Which means that: $\{Y_i(0); Y_i(1)\} \perp\!\!\!\perp T_i|\mathbf{X}_i$ for all i.

3. Stable Unit Treatment Value Assumption (or **SUTVA**). The implication of this assumption is twofold: (1) the treatment assignment of any unit does not affect the potential outcomes of other units (often referred to as noninterference) and (2) there is only one version of the treatment, this implies that the treatment is comparable across units.

4. Experimental treatment assumption or **Positivity**. This assumption implies, that for a given set of confounders, each unit has a positive probability of being assigned to the treatment group (i.e, bounded away from zero)

5. **Absence of measurement error**. This assumption implies that the observed outcome, the treatment indicator and all the confounders are measured without error.

6. Independent and identically distributed observations. Most of the causal inference methods are developed for **simple random samples**.

7. **Correct model specification**. In order to correctly estimate causal effects,

causal inference assumes that the statistical models for the outcome and for the probability of receiving treatment are correctly specified. This assumption can be relaxed when a doubly robust (DR) estimator is used. When a DR estimator is computed, it suffices to have one model correctly specified in order to obtain consistent estimators of causal effects (see Chapter 2).

Although randomized clinical trials (RCT) are considered the gold standard to evaluate treatments, situations in which randomization of a treatment/exposure is not feasible are not uncommon, especially in public health research. One of the most appealing features of the RCM, is that its conceptual framework is broad enough that allows for the estimation of causal effects using non-experimental data. Methods like instrumental variables, regression analysis, regression discontinuity, propensity score stratification, propensity score matching and propensity score weighting, among others, were promptly developed for this purpose. In this manuscript, we focus on three propensity score based methods that are commonly implemented: (1) propensity score matching, (2) propensity score weighting and (3) doubly robust estimators.

The goal of this manuscript is two-fold, on one hand we relax assumptions 5 and 6, and on the other, we explore the consequences of violating assumption 7. When working with non-experimental data, especially with self-reported information, measurement error tends to be the rule rather than the exception. Therefore, it is important to develop methods that can estimate causal effects under the presence of measurement error. By relaxing assumption 6, we extend causal inference methods to a context

of complex survey data. Large scale, complex survey designs are becoming increasingly common in public health. In such large-scale surveys, the sampling framework may be complicated, and the sampling probabilities (and resulting survey weights) vary depending on the sampling of sub-populations (these survey weights may then also be adjusted to account for non-response or to post-stratify to known population totals). When attempting to make inferences to the target population, the survey design (e.g., survey weights and design elements) should be correctly incorporated in the data analysis; otherwise, the parameter estimates may not be relevant to the original target population of the survey. Finally, by assessing the consequences of model misspecification in the propensity score and outcome model, we can evaluate which causal effect estimators perform better under different degrees of misspecification.

The rest of this manuscript is organized as follows: in Chapter 2 we extend the Simulation-Extrapolation methodology or SIMEX (Cook and Stefanski, 1994; Carroll et al., 1996) to a doubly robust estimator of the ATE. This extension allows to mitigate the bias in the estimation of the ATE, induced having a single covariate measured with error. Furthermore, we present a new structure of measurement error. This new structure is more general than the classical measurement error structure, allowing us to capture features associated with self-reporting. We illustrate the application of this method using the National Longitudinal Study of Adolescent to Adult Health (Add Health) data, to estimate the effect of depression on sexual health. In Chapter 3, we present a propensity score matching estimator of the ATT in the context of

complex survey data. In this chapter, we present guidelines on how to handle survey weights when estimating the ATT using propensity score matching and compare the performance of different estimation approaches commonly used in the literature. We illustrate the application of such guidelines by using the Early Childhood Longitudinal Study, Kindergarten class 1998-1999 (ECLS-K) to estimate the effect of special education services on math skills. In Chapter 4 we compare the impact of model misspecification when implementing two of the most commonly used propensity score based methods: propensity score matching and weighting. In this chapter, we present a measure of the degree of misspecification and evaluate the performance of the estimators considered as a function of the model misspecification. Finally in Chapter 5, we present our main conclusions and provide a summary of our main findings.

Excerpts of this chapter and Chapter 2 are from the paper entitled "A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error", which has been published in *Biostatistics* (Lenis et al., 2017a).

Chapter 2

A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error

2.1 Introduction

In many fields measurement error tends to be the rule rather than the exception. Methods such as Simulation-Extrapolation (SIMEX) (Cook and Stefanski, 1994), regression calibration (Rosner et al., 1990) and multiple imputation (Cole et al., 2006; Guo et al., 2012) have been developed to mitigate the impact of measurement

error in the estimation of coefficients, but limited work has been done to extend these approaches to a causal inference context.

Causal inference has provided a number of tools and a clear conceptual framework in which causal effects can be estimated. The development of causal methods has had an important impact in the design of clinical trials and sampling designs and the framework has been extended to observational studies and other study designs where formal randomization is not possible. In particular, since propensity-score based methods were first introduced by Rosenbaum and Rubin (1983), a wide range of methods based on propensity scores have been developed to estimate treatment effects in non-experimental studies. Methods such as matching, weighting or subclassification (Stuart, 2010) allow for comparison of treatment and control groups that are similar based on a set of observed characteristics. One can also use doubly robust estimators (Rotnitzky et al., 1998), which utilize models of treatment assignemnt (the propensity score) and of the outcome, to estimate a treatment effect. One benefit of doubly robust estimators is that they have an asymptotically normal distribution and furthermore, they are consistent if either the model for the propensity score or for the conditional mean of the outcome (but not necessarily both) are correctly specified. Nevertheless, all of these methods rely on the assumption that all covariates, treatment indicator and observed outcome are measured without error.

Steiner et al. (2011) has shown that measurement error in the covariates can lead to bias in the treatment effect estimator, when the true propensity score model depends

on the unobserved true covariates. McCaffrey et al. (2013) has also shown that when at least one of the covariates is measured with error, balance between the treatment and control groups on the true (unobserved) covariate is not always achieved. There is a need to extend and develop methodology to account for measurement error in a causal inference context. Stürmer et al. (2005) propose a propensity score calibration method to account for unmeasured confounders (i.e., the true covariate in the context of measurement error). This method is related to the regression calibration approach and relies on a validation sample. Nevertheless their method can only apply when the validation sample has information regarding the set of all relevant covariates and the treatment assignment. McCaffrey et al. (2013) propose a measurement-error bias-corrected inverse probability of treatment weighting estimator. This method requires either the distribution of the measurement error or the unobserved true covariate to be known, and the propensity score model to be correctly specified. Webb-Vargas et al. (2015) implement a multiple imputation approach to correct for covariate measurement error in propensity score estimation, and compute a doubly robust estimator of the treatment effect. However, this method requires knowledge regarding the joint distribution of the variables (covariates, outcome and treatment indicator). Furthermore, convergence problems have been reported when many binary confounders are included. Finally, Lockwood and McCaffrey (2015) extended the SIMEX methodology to causal inference in the context of typical classical measurement error (i.e., the error term is additive, non-differential and homoscedastic). In this article we show

that the SIMEX methodology can be extended to a more complex measurement error structure that has the typical classical measurement error structure as a special case.

Mean-reverting measurement error structures have been described by Bound and Krueger (1991) in the context of longitudinal data of earnings. As stated by Akee (2011), mean-reverting measurement error in the context of self-reported earnings implies that "the higher the true value of earnings, the more likely an individual is to under report her earnings and vice versa". In general terms and in the context of self reported variables, a mean-reverting measurement error implies that the units with larger values of a given variable tend to underreport such values whereas units with smaller values then to overreport their true value. Mean-reverting measurement error is traditionally modeled with a similar structure as the typical measurement error with the exception that the measurement error is negatively correlated with the true value of the mismeasured covariate. In this article we present an alternative way to model a mean-reverting measurement error. The main advantage of our proposed parametrization is two-fold: (1) it allows for "mean-diverging" measurement error (i.e. higher values of the true covariate are associated with even larger reported values and vice versa) and (2) the typical classical measurement error structure can be conceived as a special case of this more general measurement error structure.

We propose to extend the SIMEX methodology to a doubly robust estimator of the average treatment effect, when a covariate is measured with error (under a mean-reverting measurement error structure) but the treatment, outcome and the rest of

the covariates are measured without error. This method does not require assumptions regarding the joint distributions of the variables. Additionally, the validation data only needs to have information regarding the true covariate and the faulty measured version. Furthermore, the measurement error structure (i.e., reverting, diverging, or typical) does not have to be specified beforehand. Validation samples are not uncommon in studies where acquiring the true value of the covariate of interest is too expensive, time consuming, or invasive (Pettersen et al., 2012; Saint-Maurice et al., 2014). Work by Robins (2003), Cole et al. (2006), Goetghebeur and Vansteelandt (2005), or Edwards et al. (2015) has examined the consequences of measurement error in the outcome and/or the exposure.

This article is organized as follows: Section 2.2 presents definitions and working models, Section 2.3 introduces a doubly robust estimator and summarizes the SIMEX method, Section 2.4 deals with the asymptotics, Section 2.5 presents the results of a simulation study, Section 2.6 presents an application of the method using the National Longitudinal Study of Adolescent to Adult Health (Add Health), and Section 2.7 presents our conclusions.

2.2 Definitions

2.2.1 Measurement Error Structure

Different measurement error structures have been proposed in the literature and most of them can be grouped in two categories: classical and Berkson. Classical measurement error structures assume that the true value of a covariate is not observed but a faulty version of it is available (which in the literature is referred to as a "surrogate"). In contrast, Berkson measurement error happens "when a group's average is assigned to each individual suiting the group's characteristics. The group's average is thus the 'measured value', that is, the value that enters the analysis, and the individual latent value is the 'true value'." (Heid et al., 2004). Besides the technical differences between classical and Berkson type error, the main difference between these two structures is related to the consequences in the estimation of parameters. For example, in the context of linear regression, it can be shown that under classical measurement error structures regression coefficients will be inconsistent. However, under Berkson error structures the estimators, although inefficient, will be consistent.

Throughout this article we assume a measurement error structure that belongs to the classical type and that affects only one of the covariates. If X_i is defined as the true and unobserved value of a covariate for unit i, the observed surrogate measure

of X_i, say W_i, is assumed to be of the form:

$$W_i = X_i + \tau_1 \left[X_i - E\left(X_i\right)\right] + \sigma\epsilon_i \tag{2.1}$$

where $E\left(X_i\right)$ is the expected value of the mismeasured covariate. Notice that different configurations of σ and τ_1 may lead to different measurement error structures. For example, if $\tau_1 = 0$ the measurement error structure follows a typical classical measurement error structure. Furthermore we could potentially find combinations of values for τ_1 and σ such that the measurement error could be either mean reverting or mean diverging. Negative values of τ_1 are associated with mean reverting measurement error structures while positive values of τ_1 are associated with mean diverging structures. Observe that if $\tau_1 = -1$, W represents random deviations from the mean of X. Also notice that for a given value of σ and depending on the value of τ_1, it is possible that the variance of the surrogate W will be smaller than the variance of the true covariate X. Therefore, the notion of reliability $\left(\text{i.e. } \frac{Var(X)}{Var(W)}\right)$ is no longer fully informative. Under this measurement error parametrization $\sigma\epsilon_i$ represents the difference in the reported values among units with the same true value of the covariate X. We assume that ϵ_i is a random variable following a normal distribution with mean zero and unit variance that is independent of X_i for all i.

2.2.2 Working Models

The goal of this article is to estimate the average treatment effect (ATE) of a binary treatment (T) on an observed outcome (Y, with $Y \in \mathbb{R}$) when a set of covariates (X, Z) are available (with $X \in \mathbb{R}$ and $Z \in \mathbb{R}^q$).

2.2.2.1 Propensity Score

We define the propensity score for unit i as the probability of receiving treatment given the covariates (X, Z). Explicitly: $P(T_i = 1|X_i, Z_i) = \pi_i$ where:

$$\pi_i(\cdot) \in \left\{\pi(X_i, Z_i; \alpha) : \alpha \in \mathbb{R}^{q+2}\right\} \tag{2.2}$$

and $\pi(\cdot) : \mathbb{R}^{q+1} \to (0, 1)$ is a parametric model that includes an intercept. We also assume the first derivatives are defined, namely: $D_{\alpha,i}^T = \frac{\partial \pi_i}{\partial \alpha}^T = \left[\begin{array}{ccc} \frac{\partial \pi_i}{\partial \alpha_1} & \cdots & \frac{\partial \pi_i}{\partial \alpha_{q+2}} \end{array} \right]^T$.

2.2.2.2 Conditional Mean model

We assume the conditional mean model to be:

$$E[Y_i|X_i, Z_i, T_i] = \mu_i \;\; = \;\; \beta_0 + \Delta T_i + X_i\beta_X + Z_i^T\beta_Z \tag{2.3}$$

with β_0, β_X and $\Delta \in \mathbb{R}$ and $\beta_Z \in \mathbb{R}^q$. We define: $D_{\beta_0,i} = \frac{\partial \mu_i}{\partial \beta_0}$; $D_{\beta_X,i} = \frac{\partial \mu_i}{\partial \beta_X}$; $D_{\beta_Z,i}^T = \left[\begin{array}{ccc} \frac{\partial \mu_i}{\partial \beta_{Z,1}} & \cdots & \frac{\partial \mu_i}{\partial \beta_{Z,q}} \end{array} \right]^T$; $D_{\Delta,i} = \frac{\partial \mu_i}{\partial \xi}$. We let $D_i^T = [D_{\beta_0,i}, D_{\beta_X,i}, D_{\beta_Z,i} D_{\xi,i}]^T$, notice that

under this model specification, and under the assumptions and regularity conditions described in Abadie and Imbens (2016) the ATE is equal to Δ.

2.3 Doubly Robust Estimators and SIMEX

2.3.1 Doubly Robust Estimators

Under regularity conditions, if (2.2) or (2.3) are correctly specified, then it can be shown that there exists a consistent and normally distributed estimator for $\Theta_0 = \left\{\beta_0, \beta_X, \beta_Z^T, \Delta, \alpha^T\right\}^T$,with $\Theta_0 \in \mathbb{R}^{2q+5}$ (Rotnitzky et al., 1998). Doubly robust estimators can be formulated in the context of estimating equations (Robins et al., 2007). Define $\omega_i = \frac{T_i}{\pi_i} + \frac{1-T_i}{1-\pi_i}$; and let $\Gamma_i^T = D_i^T \frac{1}{v_i} \omega_i \left[Y_i - \mu_i\right]$. We now define the following vector: $\psi_1 \left(X_i, Z_i, T_i, Y_i, \Theta\right) = \left[\Gamma_i, D_{\alpha,i}, \Delta - E\left[Y_i | X_i, Z_i, T_i = 1\right] - E\left[Y_i | X_i, Z_i, T_i = 0\right]\right]^T$ where Θ is a vector of parameters in $\mathbb{R}^{2q+6}$. If the propensity score or the conditional mean model are correctly specified then $E\left[\psi_1\right] = 0$. Thus if we define $\widehat{\Theta}_{DR}$, such that $\sum_{i=1}^n \psi_1 \left(X_i, Z_i, T_i, Y_i, \widehat{\Theta}_{DR}\right) = 0$ then $\widehat{\Theta}_{DR}$ will be a consistent estimator for Θ_0 and will have an asymptotically normal distribution. Additionally, we assume that if both models are incorrectly specified, the resulting estimator will converge to Θ^* which is not necessarily equal to Θ_0 and the estimator will follow an asymptotically normal distribution. Explicitly, if we define $\tilde{\Theta}$ to be the solution to $\sum_{i=1}^n \psi_1 \left(X_i, Z_i, T_i, Y_i, \tilde{\Theta}\right) = 0$ then we can conclude that $\tilde{\Theta} \to \Theta^*$, where Θ^* may be different from the true vector of parameters Θ_0.

2.3.2 SIMEX

From this point forward, we assume that the propensity score is correctly specified. Given that we are implementing a doubly robust estimator of the ATE, if our method is able to account for the impact of the measurement error in the prediction of the propensity score the final SIMEX estimator of the treatment effect will be consistent even when the model for the conditional mean of the observed outcome is misspecified. Therefore, we can estimate the vector of unknown true parameters Θ_0 with $\hat{\Theta}$, by solving the following unbiased estimating equations: $\sum_{i=1}^{n} \psi_1 \left(X_i, Z_i, T_i, Y_i, \widehat{\Theta}_{DR} \right) = 0$. Since the variable X is not observed, the solution to $\sum_{i=1}^{n} \psi_1 \left(W_i, Z_i, T_i, Y_i, \widehat{\Theta}_{NA\ddot{I}VE} \right) = 0$ will lead to an inconsistent estimator of the vector of parameters of interest. Therefore the solution to the estimating equations is a consistent estimator for some other vector of parameters, say Θ^*. This property of convergence to some vector of parameters is fundamental for the implementation of the SIMEX methodology. Since convergence is always achieved (even in the presence of measurement error), we can artificially increase the measurement error in the surrogate W, and evaluate the trend of the bias as a function of such increments. Then we can extrapolate to the case of no measurement error. We now describe the two steps in SIMEX: simulation and extrapolation.

2.3.2.1 Simulation Step

Let B be a large fixed positive integer. Then for each unit i we generate B random standard normal variables, ϵ_{ib}, indexed by $b = 1, 2, ..., B$. We define W_{ib} as $W_{ib} = W_i + \sqrt{\lambda}\sigma\epsilon_{ib}$ for some fixed $\lambda > 0$ (for simplicity, σ is assumed to be known) . Let $\hat{\Theta}_{b,\lambda}$be the solution to $\sum_{i=1}^{n} \psi_1 \left(W_{ib}, Z_i, T_i, Y_i, \hat{\Theta}_{b,\lambda}\right) = 0$. Then we can define $\hat{\Theta}_\lambda = \frac{1}{B}\sum_{i=1}^{B} \hat{\Theta}_{b,\lambda}$. Now we can repeat this procedure for $\Lambda = \{\lambda_k : \lambda_k > 0 \; for \; k = 1, 2, 3..., K\}$ to obtain $\left\{\lambda, \hat{\Theta}_\lambda\right\}_{\lambda\in\Lambda}$. This sequence allows us to evaluate the trend of the bias as a function of the increments in the measurement error and extrapolate to the case of no measurement error, when $\lambda = -1$.

2.3.2.2 Extrapolation Step

Cook and Stefanski (1994) proposed to compute the SIMEX estimator as $\hat{\Theta}_{SIMEX} = \mathcal{G}\left(\hat{\vartheta}, -1\right)$, where $\mathcal{G}(\vartheta, \lambda) : \mathbb{R} \to \mathbb{R}^{2q+6}$ is a parametric model for the vectors $\hat{\Theta}_\lambda$ as a function of Λ and $\vartheta \in \mathbb{R}^p$ is a p-dimensional vector of coefficients associated with the model. If at least one but, not necessarily both of the working models (i.e., the propensity score or the conditional mean model) are correctly specified and the parametric model $\mathcal{G}(\cdot, \cdot)$ is also correctly specified, then $\hat{\Theta}_{SIMEX}$ will be a consistent estimator of the true vector of parameters Θ_0 (Carroll et al., 1996).

2.4 Asymptotics

When Cook and Stefanski (1994) presented SIMEX, they suggested a bootstrap procedure to compute standard errors of the SIMEX estimator. A few years later, Carroll et al. (1996) derived the asymptotic distribution of the SIMEX estimator under a typical classical measurement error structure and the assumption that σ is know, Grace (2008) extended these results to longitudinal data. Carroll et al. (1996) provided guidelines to estimate the distribution of the SIMEX estimator when the variance of the measurement error is unknown but it can be estimated with an asymptotic normal estimator. Furthermore, following closely the derivation presented by Carroll et al. (1996) a valid asymptotic distribution of the ATE can be derived even when the measurement error has the structure presented in equation 2.1. Thus, we only need to specify a valid estimator for the variance of the measurement error when a validation sample is available. Notice that in the validation sample, both X and W are observed. We denote with m the sample size of the calibration sample and express equation 2.1 as $W_i = -\tau_1 E(X_i) + (1+\tau_1)X_i + \sigma_\epsilon \epsilon_i$ for $j = 1, 2, ...m$. Therefore the variance of the measurement error can be estimated by the sample variance of the residuals of the simple linear regression of X on W. In other words, the estimator of σ^2 can be expressed as if $\widehat{\sigma^2} = \frac{1}{m}\sum_{j=1}^{m} e_j^2$ where e_j represents the j^{th} residual of the simple linear regression of X on W. It can be shown that $\sqrt{m}\left(\widehat{\sigma^2} - \sigma^2\right) \xrightarrow{D} N(0, \Omega)$ where Ω can be computed using influence functions. More explicitly $\Omega = E[\varphi^2]$ with $\varphi = (W^2 - E[W^2]) - 2(1+\tau_1)(WX - E[WX]) - 2\tau_1 E(X)(1+\tau_1)(X - E[X]) +$

$(1 + \tau_1)(X^2 - E[X^2])$.

It is important to note that the measurement error structure presented in Section 2.1 is not innocuous: it can be shown that even after applying the SIMEX methodology, when the measurement error has the structure defined in equation 2.1, the estimator of the coefficient associated with the covariate measured with error will be inconsistent. A motivating example showing this and a procedure to obtain a consistent estimator of the coefficient associated with the missmeasured variable are available in Appendix A in Supplementary Materials.

2.5 Simulation Study.

To evaluate the performance of our estimator we conduct a simulation study to compare bias, mean squared error (MSE) and coverage of three different estimators of the treatment effect Δ: (1) the estimator obtained by using X, the true measure of the covariate, (2) a naive estimator, which ignores the measurement error and simply uses W, and (3) the SIMEX estimator for the treatment effect. The three methods implement a doubly robust approach using propensity score weights. A total of 1000 simulation iteration were implemented. We set $\mathcal{G}(\vartheta, \lambda)$ as a quadratic function; explicitly $E\left(\hat{\Theta}_\lambda|\lambda\right) = \omega_0 + \omega_1\lambda + \omega_2\lambda^2$. Details of the data generating process can be found in Appendix B in Supplementary Materials. In the simulation study, we fit a correctly specified propensity score model, but we fit the following

model for the conditional mean: $E(Y_i|T_i) = \eta_0 + \eta_1 T_i$. Notice that we have purposely omitted Z and X, thus incorrectly specifying the conditional mean model. Since we are using a doubly robust estimator, it holds that $\widehat{\eta_1}$ will converge to Δ if our method is able to account for the impact of the measurement error in the prediction of the propensity score. We evaluate the performance of the SIMEX estimator described in Section 3.2 and compare it to that of the naive estimator and the estimator obtained when the covariate measured without error is used. Figure 1 summarizes our main findings.

As expected, when the true covariate is used in the estimation, performance is very good and is used as the baseline for comparisons. The naive method (ignoring the measurement error) leads to biases in the estimated treatment effect across all the settings considered. Furthermore, the bias decreases as τ_1 increases. In terms of bias, the SIMEX estimator outperforms the naive method, and this result holds for all correlation levels of X and Z, and across the different coefficients on X in the true treatment assignment (propensity score) model. Similar patterns observed for MSE and coverage, in terms of the SIMEX estimator performing better than the naive approach.

In general, we observe that the SIMEX approach performs better when the coefficient on the true covariate in the propensity score model is small and when the correlation between the covariates is relatively low. Notice that all methods can produce coverage above 95%. This is due to the fact that the estimated propensity score

is used in the computation of the estimators' weights, and thus the standard errors are overestimated (Rubin and Thomas, 1996; Rubin and Stuart, 2006). Notice that the same conclusions hold even when $\tau_1 = 0$ (i.e., when the measurement error follows a typical classical structure) which implies that the defined measurement error structure defined in Section 2, can easily accommodate for a typical measurement error structure. In general we observe that, as expected, the estimator that uses X as a regressor performs better than the SIMEX and the Naive estimators across all simulated scenarios. The performance of the Naive method suggests that ignoring the measurement error in a covariate, can induce bias in the estimated treatment effect which translates into higher MSE and poor coverage. This situation is exacerbated when the impact of X in the propensity score is large and the correlation, ρ, between X and Z is high. The simulation study suggests that, implementing the SIMEX methodology can help to mitigate the consequences associated with measurement error.

2.6 Application

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a multi-year longitudinal study of a nationally-representative sample of adolescents in the United States that began during the 1994-95 school year, when the adolescents were in grades 7-12. Information regarding a wide range of topics (e.g., socioeconomic

factors, relationships, psychological and physical health, etc.) was collected during four waves. For details see Harris et al. (2009). In this application, we estimate the effect of depression (the exposure) on sexual health, where Body Mass Index (BMI) is the confounder measured with error.

We use the Add Health data to evaluate the performance of the SIMEX estimator in a realistic data context. For this application, we use the publicly available Add Health data of subjects who participated in Waves I and II. This dataset present a unique feature: during the second wave BMI was both measured and self-reported. Thus we can compute the treatment effect using the true BMI, and compare the result to those obtained implementing SIMEX and those obtained using the naive approach (using the self-reported BMI). To do this, we artificially construct a validation sample by randomly selecting $\frac{1}{6}$ of the observations (this is the same relative sample sizes used in simulation study). The variance ratio of the self-reported BMI to the measured BMI is equal to 0.90 which indicates that the measurement error structure cannot follow the typical classical structure, since under that structure the variance of the surrogate is always larger than the variance of the true covariate. Furthermore, the correlation between the self-reported and the measured BMI is 0.92 and the R^2 associated with a simple linear regression of the self-reported measure on the measured BMI is 0.84. This indicates that the self-reported BMI is a highly reliable measure of the true BMI and therefore we do not expect significant differences between the different treatment effect estimations. In fact, the estimated

treatment effect was 0.052 and statistically insignificant regardless of the approach used to estimate it.

Thus, we propose a data-based simulation study where all the covariates are obtained from the Add Health data, but the outcome, the exposure and the variable measured with error are simulated. By controlling the data generating process we should be able to assess the performance of the SIMEX estimator in more complex data structure.

2.6.1 Data-based simulation set-up

The Add Health data contains the measured weight and height of all the adolescents in Wave II, and so a highly reliable measure of the Body Mass Index (BMI) can be obtained. Plankey et al. (1997) and Stommel and Schoenborn (2009) model self-reported BMI in the context of mean reverting measurement error. In order to evaluate the performance of the SIMEX estimator, we construct a self-reported BMI, $srBMI$, as $srBMI_i = BMI_i + (1 - \tau_1)(BMI_i - E[BMI_i]) + \sigma\varepsilon_i$ with independently and identically distributed errors $\varepsilon_i \sim N(0, 1)$ and variance equal to $\sigma^2 = 0.3 \times Var(BMI_i)$. Both the $Var(BMI_i)$, the $E[BMI_i]$ and τ_1 are estimated from the available data.

The set of covariates measured without error, Z, are listed in Table 3.1. These variables have been suggested by Goodman and Whitaker (2002) to have an effect on depression, which constitutes the exposure. Goodman and Whitaker (2002)

also link depression with BMI. Thus, we construct an indicator of depression status, $mdep1$, assuming that $mdep1_i \sim Bernoulli\,(p_i)$, with $p_i = expit\,(-3.3 + \alpha BMI_i + 1.1NotWhite_i + 6welfare1p_i - .5PEd\_CC_i + .5delinqI_i + .8heavyds_i)$.

All coefficients, except the one associated with BMI, are chosen based on a estimated logistic regression using the available data. In section 2.5 we have shown that a strong association between the true values of the unobserved covariate (BMI) and the exposure (depression) affects the performance of the SIMEX estimator, in other words the stronger the association between the missmeasured covariate and the exposure, the more compromised is the performance (in terms of bias, coverage and MSE) of the SIMEX estimator. Thus we increased the association between BMI and depression by a factor of twenty, which implies that α (the coefficient on BMI in the propensity score model) is equal to 0.066.

Wingood et al. (2002) suggests that BMI and depression affect sexual health. We generate the outcome variable, number of different sexual partners in the last year, $npartneryear$, from a normal distribution. That is $npartneryear_i \sim N\,(\mu_i, 1.4^2)$, with: $\mu_i = \beta + \beta BMI_i + \beta mdep1_i - \beta NotWhite_i + \beta welfare1p_i - \beta PEd\_CC_i + \beta delinqI_i + .\beta heavyds_i$. For simplicity we set $\beta = 1.5$. Out of the 2640 complete cases we randomly select 440 adolescents that will constitute the validation sample (i.e., the sample where BMI and the generated variable measured with error, $srBMI$, are observed). The remaining 2200 observations constitute the main sample, the sample where BMI is not observed, but its surrogate is. The doubly robust estimator is

computed using the data from the main sample. We ran a total of 1,000 iterations

2.6.2 Data-based simulation results

The main results from the data-based simulation are summarized in Table 2.2 . The first column of Table 2.2 shows the name of the covariates used in the outcome model, in the second column the true value of the estimated parameters are displayed, in the third column we computed the average value of the estimated parameter (across the 1,000 iterations), the fourth column gives the percentage bias (in absolute value), the fifth column shows the empirical coverage of the 95% confidence interval and finally, the last column gives the MSE. Part I of Table 2.2 shows the estimation results using the measured BMI. As expected, the bias is negligible and the empirical coverage confidence intervals is close to 95% for all the covariates included in the outcome model. Part III of Table 2.2 shows the estimating results associated with the naive method (i.e., using the generated self-reported BMI), the performance of the naive estimator is far from ideal and the estimators of the coefficients associated with the variables $mdep1$, $NotWhite$ and $delinqI$ have on average biases larger than 10%. This is particularly important in the estimated treatment effect (i.e., the coefficient associated with the variable $mdep1$) where the bias is about 32%. Part II of Table 2.2 presents the estimation results obtained by implementing the SIMEX method. On average almost all of the coefficients have biases less than 5% (the only exception is the estimator associated with the covariate $mdep1$ that has a bias of 5.01%).

The results shown in Table 2.2 incorporate the correction suggested in Appendix A in Supplementary Materials. The bias of the SIMEX estimator associated with the estimation of the treatment effect is 5.01%, in other words SIMEX was able to remove about 84% of the bias associated with the naive estimation of the treatment effect. It is important to notice that the standard errors associated with the SIMEX estimators tend to be larger that the ones obtained by the other two methods. This could potentially translate into a power loss, nevertheless the comparison of the MSE of the SIMEX estimators to that of the naive approach, suggests that the efficiency loss is negligible.

2.7 Conclusions

In this article we propose a new structure of measurement error that has the typical classical measurement error structure as a special case. We found that using a covariate measured with error can lead to biases in the estimation of the average treatment effect in non-experimental studies even when a doubly robust estimator is utilized. Our theoretical results and simulation study suggests that the SIMEX estimator can help to mitigate this problem, and a data-based simulation suggests that the SIMEX estimator can help to reduce up to 84% of the bias introduced in the estimation of the treatment effect using the covariate measured with error. It is important to highlight that the SIMEX estimator also helps to reduce the bias of the

other estimated coefficients in the outcome model.

Compared to other methods that address measurement error in a causal framework (Stürmer et al., 2005), our use of SIMEX only requires information information related to the covariate and its surrogate in the validation sample. The data-based simulations suggest that this methodology can be applied to complex data structures with multiple binary covariates, which is an improvement over the Multiple Imputation for External Calibration approach (Webb-Vargas et al., 2015), which assumes joint multivariate normality of the covariates. Future work should further investigate the relative performance of these methods under a wider range of settings.

The main limitation of the SIMEX approach is the assumption that the parametric model $\mathcal{G}(\vartheta, \lambda)$ is correctly specified. This assumption is not testable, and future work should investigate how robust the SIMEX estimator is to different model specifications (Cook and Stefanski, 1994). In addition, the method presented in this article only considers the case of a linear outcome model; further work will concentrate on extending this approach to different parametrizations, such as general linear models.

In conclusion, we have shown that estimating an average treatment effect using a doubly robust estimator in non-experimental studies can lead to significant biases when a mismeasured covariate is used in the estimation. However, the SIMEX estimator can be used to mitigate this problem. This extension is particularly relevant to public health research, where measurement error tends to be the rule rather than the exception.

2.8 Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors thank Elizabeth M. Sweeney and John Muschelli for comments that greatly improved this manuscript. This work was supported by the National Institute of Mental Health (R01MH099010; PI: Elizabeth A. Stuart.) *Conflict of Interest*: None declared.

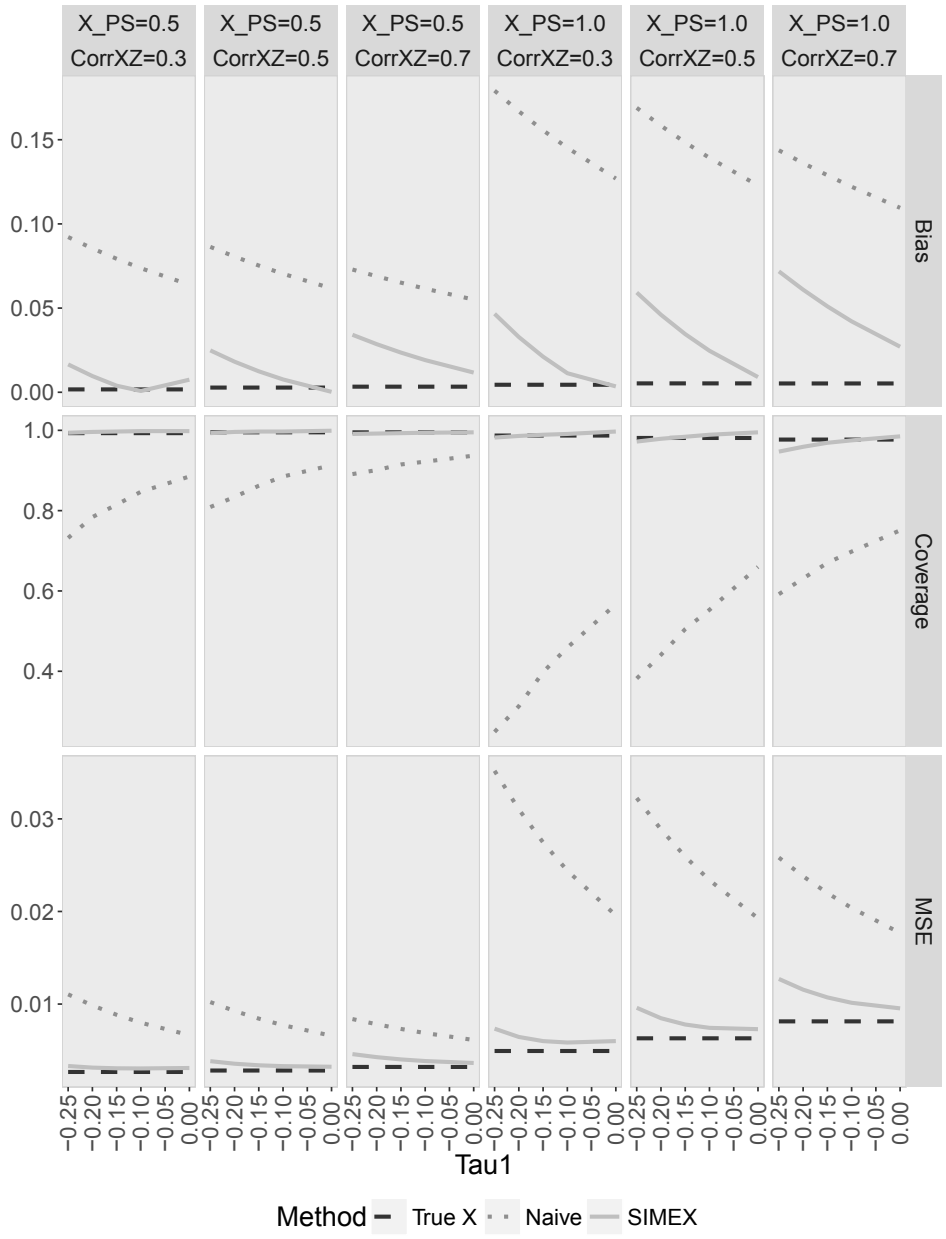

Figure 2.1: Absolute bias, coverage and mean squared error (MSE) as functions of τ_1 for different levels of correlation of the covariates and effect size of the unobserved variable in the propensity score (simulation study).

Table 2.1: Empirical example using Add Health data: Covariates used in propensity score and outcome models

Wave	Covariates	Code	Description
I	Race	$NotWhite$	Indicator variable indicating if the respondent did not answer that his or her race is White
I	Welfare	$welfare1p$	Indicator variable indicating that at least one of the subject's parents responded that he or she receives government assistance.
I	Parent's Education	PEd_CC	Indicator variable that identifies if at least one parent completed college or higher level education.
II	Heavy Drinking or Smoking	$heavyds$	Indicator variable that takes the value one if the adolescent is either a heavy smoker or a heavy drinker or both. Heavy smokers are those individuals for whom the amount of cigarettes smoked in the last month is in the top quartile. Heavy drinkers are those individuals who have had at least three drinks per week. These cutoffs are based on those used in Goodman and Whitaker (2002).
II	Delinquent Behavior	$delinqI$	Indicator variable that identifies if the adolescent was seriously involved in criminal activities, as measured by scoring in the top quartile of the Add Health delinquency scale. This dichotomization follows the same criteria as Goodman and Whitaker (2002).

Table 2.2: Estimation results from the data-based simulation.

	Part I True			
	Average	**Bias**	**Coverage**	**MSE**
Depression ($mdep1$)	1.50	0.21%	94.8%	0.006
Race ($NotWhite$)	-1.50	0.06%	95.5%	0.007
Welfare ($welfare1p$)	1.51	0.39%	94.1%	0.016
Parents' Education (PEd_CC)	-1.50	0.10%	93.6%	0.010
BMI	1.50	0.00%	94.2%	0.000
Delinquent Behavior ($delinqI$)	1.51	0.34%	94.1%	0.009
Heavy Drinking or Smoking ($heavyds$)	1.50	0.06%	95.4%	0.008
	Part II SIMEX			
	Average	**Bias**	**Coverage**	**MSE**
Depression ($mdep1$)	1.42	5.01%	93.9%	0.006
Race ($NotWhite$)	-1.52	1.44%	96.1%	0.007
Welfare ($welfare1p$)	1.48	1.55%	96.5%	0.016
Parents' Education (PEd_CC)	-1.47	1.78%	94.4%	0.010
BMI	1.55	3.29%	100%	0.000
Delinquent Behavior ($delinq1$)	1.52	1.47%	95.7%	0.009
Heavy Drinking or Smoking ($heavyds$)	1.49	0.62%	96.9%	0.008
	Part III Naive			
	Average	**Bias**	**Coverage**	**MSE**
Depression ($mdep1$)	1.98	31.72%	11.9%	0.249
Race ($NotWhite$)	-1.32	11.8%	81.4%	0.053
Welfare ($welfare1p$)	1.57	4.63%	95.7%	0.055
Parents' Education (PEd_CC)	-1.64	9.32%	87.9%	0.051
BMI	1.38	7.94%	0.0%	0.014
Delinquent Behavior ($delinqI$)	1.31	12.96%	81.1%	0.061
Heavy Drinking or Smoking ($heavyds$)	1.59	5.86%	94.2%	0.032

Chapter 3

It's all about balance: propensity score matching in the context of complex survey data

3.1 Introduction

3.1.1 Background

Non-experimental data are increasingly used to estimate the causal effects of certain exposure/intervention (hereafter, 'treatment'), especially when a randomized trial is infeasible or unethical. More often than not, the interest is in causal effect estimates that generalize to an entire target population, as opposed to apply to only

a data sample. These interests combined call for the use of data that help inform about the target population and statistical methods that ensure the inferences are as accurate as possible. Two tools available for these purposes are large-scale nationally-representative datasets and propensity score methods.

Large scale, complex survey designs are widely used and usually have a well-defined target population and sampling framework. The sampling framework may be complicated, and the sampling probabilities vary depending on the sampling of sub-populations. When attempting to make inferences to the target population, the survey survey weights and other design elements should be correctly used in data analysis; otherwise, the parameter estimates may not be relevant to the original target population of the survey (e.g., Hansen et al., 1983; Korn and Graubard, 1995a; Korn and Graubard, 1995b and Little (2003)).

The causal inference framework introduced by Rubin (1974) extended the estimation of causal effects to non-experimental studies. Since propensity scores (i.e., the probability of receiving treatment given a set of observed covariates) were introduced by Rosenbaum and Rubin (1983), a wide range of methods have been developed to estimate treatment effects in non-experimental studies (e.g., propensity score based matching, weighting, and subclassification).

In particular, propensity score matching estimators have been widely used in the context of non-experimental studies. Matching methods help reduce bias in the estimation of causal effects (see Rubin, 1973a), and are intuitive and relatively easy to

implement. However, standard propensity score matching methods do not give guidance on how to incorporate survey weights, and conceptually it is somewhat unclear how to do so. As a consequence, researchers using propensity score matching methods often do not incorporate the complex survey design (e.g., Morgan et al., 2008). This paper aims to provide guidance on propensity score matching using complex survey data to ensure that the estimated causal effects apply to the target population.

3.1.2 Previous Research in this Area

There has been extensive work in each of the two areas to be investigated in this article (complex surveys and propensity scores), but only limited work on how to combine them.

Propensity score methods have been developed under the assumption of a simple random sample (SRS), yet this sampling scheme is hardly ever used since every unit in the population has to be listed, making this sampling method very cumbersome to use for large populations. To guarantee representation of the population, complex survey techniques such as stratification and clustering may be implemented. In addition to the sampling design, certain adjustments (e.g., adjustment for non-response or post-stratification to match population composition) are also built into survey weights, which are used to scale the sample back to the population.

There is a general consensus that ignoring survey weights leads to external validity bias, because inferences about the population are based on a unrepresentative analytic

sample. Thus, survey weights and the sampling design should be incorporated in the estimation process. It has been widely documented how to incorporate survey weights in the estimation of means, totals and ratios (see Cochran, 1977; Groves et al., 2009), nonetheless there is controversy over how to incorporate survey weights in more complex statistical analysis (see Gelman, 2007), and propensity score methods are no exception to that.

A propensity score based analysis includes two key stages: (1) estimating propensity scores, and (2) using them in the estimation of causal effects. Regarding whether to use the survey weights in the estimation of the propensity scores, Brunell and DiNardo (2004) and Heckman and Todd (2009) argue that it is fine to not do so, because "the odds ratio of the propensity score estimated using misspecified weights is monotonically related to the odds ratio of the true propensity scores" (Heckman and Todd (2009) p.3), "and therefore does not change the relative weighting of the data" (Brunell and DiNardo (2004), p.32). Ridgeway et al. (2015) argue – in the context of propensity score weighting – that survey weights should be incorporated in the estimation of the propensity scores, and failure to do so may lead to inconsistent estimators. Austin et al. (2016) explore propensity score matching and conclude that whether survey weights are incorporated in propensity scores estimation does not affect the performance of the causal effects estimators.

There are also questions about how the survey weights should be used in the second stage: the use of the propensity scores – for example, after implementing a propensity

score matching procedure, whether survey weights need to be incorporated to assess the balance of the covariates, or, in the context of propensity score weighting, how the final weights should be constructed. There is thus a broad array of approaches used in the literature, and until recently there was almost no methodological work on the best ways to use propensity scores with complex surveys, with the exception of Zanutto (2006), Ridgeway et al. (2015) and Austin et al. (2016).

In this article we extend the scope of the analysis of previous research to incorporate different non-response mechanisms. This allows us to (1) evaluate the performance of different matching estimators in realistic scenarios (given that non-response is nearly always present), and (2) identify features of the non-response mechanism that may impact the performance of the matching estimators. As such, the main goal of this article is to identify ways in which the survey weights should be incorporated when using propensity score matching to estimate causal effects, under a variety of non-response mechanisms. **?** explored how to combine propensity score weighting and multiple imputation to correct for biases due to non-response.

The rest of this article is organized as follows: in Section 3.2 we discuss the definitions and assumptions involved in the estimation of the average causal effect, and how survey weights should be incorporated in the estimation procedure. Section 3.3 describes a simulation study and summarizes our main findings. Section 3.4 compares the performance of the different estimation procedures in an application using the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ESCL-K). In

Section 3.5 we present our main conclusions and discussion.

3.2 Definitions, Assumptions, Propensity Score and Survey Weights

3.2.1 Definitions and Assumptions

The Causal Inference Framework

Traditionally, causal treatment effects are defined based on the Rubin Causal Model (RCM) (Rubin, 1974). In the RCM, the causal effect associated with a binary treatment T (with value 1 representing the treatment of interest and 0 otherwise), is defined in terms of potential outcomes. For each unit i, $Y_i(t)$ with $t = 0, 1$ represents the outcome that would have been observed if unit i received the treatment t. For any unit i, only one potential outcome in the pair $\{Y_i(0), Y_i(1)\}$ is observed, and thus the observed outcome is $Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i)$. Given that unit level treatment effects are not identified, we are often interested in average treatment effects. At the population level, the most commonly used average treatment effects are: (1) the population average treatment effect (PATE) and (2) the population average treatment effect on the treated (PATT).

The PATE is defined as the average of the individual treatment effects over the

population, $PATE = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$, where N represents the population size. The PATT is the average of the individual treatment effects over the units in the population who were actually treated, $PATT = \frac{1}{\sum_{i=1}^{N} T_i}\sum_{i=1}^{N} T_i\,[Y_i(1) - Y_i(0)]$. When treatment effects are homogeneous (i.e., they are the same for all units in the population), the PATE is equal to the PATT. When treatment effects are not homogeneous, the PATT and the PATE can be quite different. Under of randomization of the treatment, estimation of causal effects is straightforward. Nevertheless in non-experimental data, there are a number of assumptions needed to interpret results as causal (see Rosenbaum and Rubin (1983) and Hernan and Robins (2017)), including, perhaps most importantly, that there are no unmeasured confounders, known as unconfounded treatment assignment (see, Hernan and Robins (2017))

3.2.2 Population vs. Sample Treatment Effects

Ideally, we would like to estimate population causal effects but it is rare to have full data on an entire population. In reality, causal effects are often estimated using a sample drawn from the population. Thus, we need to differentiate the PATE (PATT) from the sample ATE (ATT) – hereafter, SATE (SATT) – which is the average of the individual treatment effects for all treated units in the sample. When does a valid estimator for the SATE (SATT) correctly estimate the PATE (PATT)? The answer to this question depends on two key factors: (1) the sampling design and (2) the non-response mechanism.

Under an heterogeneous treatment effect, an unbiased estimator of the SATE (SATT) will accurately estimate the PATE (PATT) only when the sample distribution of the confounders is similar to its population counterpart. Therefore, unless survey weights are used to weight the sample back to the population, using the survey sample to estimate an ATE (ATT) will result in a consistent estimator for the SATE (SATT) but not for the PATE (PATT).

In addition, the nature of the non-response mechanism can potentially impact the estimation of the PATE (PATT). Non-response, a phenomenon by which data cannot be collected for some units that were initially selected to be in the survey sample, tends to be rule rather the exception in complex surveys. Non-response is a form of missing data. Traditionally, missing data mechanisms are grouped in three categories: (1) Missing Completely at Random (MCAR), (2) Missing and Random (MAR) and (3) Missing not at Random (MNAR) (see Little and Rubin (1989)).

Even if the sampling is designed to allow using the SATE (SATT) to estimate the PATE (PATT) (for example, SRS), if the non-response mechanism is either MAR or MNAR, the SATE (SATT) for the resulting sample may differ substantially from the PATE (PATT). Since survey weights generally incorporate non-response adjustment, failure to include them in the estimation procedure may result in misleading estimation results. More details on non-response mechanisms are available in Appendix A in the supplementary material.

3.2.3 Survey Weights and the Propensity Score

In this section we formalize the the non-response mechanisms and the propensity score model. Consider a binary indicator S_i that takes the value 1 if the i^{th} unit has been selected into the survey sample and 0 otherwise. Additionally consider a response indicator, R_i, which takes the value 1 if unit i responds to the survey. Lastly, consider $\mathbf{X}_i$ which represents a $q-$dimensional vector, for unit i, that contains all the confounders (i.e., $\mathbf{X}$ has all the covariates that are related to the treatment assignment and the potential outcomes). We assume that at the population level each $O_i = (\mathbf{X}_i, T_i, Y_i, S_i, R_i)$ is independent and identically distributed with a joint density function $f : \mathbb{R}^{q+1} \times \{0,1\}^3 \to \mathbb{R}^+$. We represent the marginal distribution for a subset of covariates $\mathbf{Z}$ (i.e., $\mathbf{Z} \subset \mathbf{X}$) with $f_{\mathbf{Z}}$. We assume that the survey sample has finite size of $n = \sum_{i=1}^{N} SR_i$, where N represents the population size, and for every $i = 1, \ldots, n$, $SR_i = S_i \times R_i$. Notice that SR_i constitutes a indicator variable that takes the value 1 if the sample unit i is selected into the survey **and** responds to the survey. We consider the case where the probability of being observed in the sample (i.e., $SR = 1$) is function of $\mathbf{X}$ and potentially the treatment indicator (T). Explicitly we assume that $p = f_{SR|\mathbf{X},T}\left(SR = 1|\mathbf{X} = \mathbf{x}, T = t\right)$ where $f_{SR|\mathbf{X},T} : \mathbb{R}^{q+1} \to (0,1)$. We assume that $p \in (0,1)$, i.e., there is not a set of values of the covariates $\mathbf{X}$ and T for which the probability of being in the sample is exactly 1 or exactly 0. Furthermore we assume that the final survey weights, ω, are equal to the inverse of the probability of being observed in the sample, that is $\omega = \dfrac{1}{p} = \dfrac{1}{f_{SR|\mathbf{X},T}\left(SR = 1|\mathbf{X} = \mathbf{x}, T = t\right)}$. These final

survey weights combine the original sampling weights (associated with the designed sampling probabilities) with corrections for non-response (see Appendix B).

We define the propensity score as $\pi = f_{T|\mathbf{X}}(T = 1|\mathbf{X} = \mathbf{x})$, the probability of receiving treatment conditional on $\mathbf{X}$, with $f_{T|\mathbf{X}} : \mathbb{R}^q \to (0, 1)$. Note that π represents the probability of receiving treatment in the population. In order to estimate π, survey weights need to be incorporated in the estimation procedure. Failure to do so will result in the estimation of the propensity score in the sample, in other words, $\pi^S = f_{T|\mathbf{X},RS}(T = 1|\mathbf{X} = \mathbf{x}, RS = 1)$, with $f_{T|\mathbf{X},RS} : \mathbb{R}^{q+1} \to (0, 1)$. Notice that if the sample distribution of $\mathbf{X}$ is different from its population counterpart, then $\pi \neq \pi^S$.

3.2.4 Survey Weights After Matching

Throughout this article, we focus on estimating the PATT. We argue that in order to estimate the PATT, survey weights may not need to be incorporated in the estimation of the propensity score model, and show that the weights of the treated units should be transferred to the comparison units to which they have been matched to, before estimating the outcome model – as suggested by Reardon et al. (2009). To see this, consider the following strategy: in a first step we implement a matching procedure using the predicted propensity score (either the $\widehat{\pi^S}$ or $\widehat{\pi}$ can be used in the procedure). We assume that k comparison units were matched without replacement to each treated observation. Now, in order to identify the weights for the treated (ω^t) and comparison units (ω^c) to use in the outcome analysis, we note that under a

successful implementation of the matching procedure, for every $\mathbf{x}$ in $\mathbf{X}$, the following equations hold:

$$f_{\mathbf{X}|T}(\mathbf{X}=\mathbf{x}|T=1) = w^c(\mathbf{x}) \times f_{\mathbf{x}|(T,M)}(\mathbf{X}=\mathbf{x}|T=1,M=1) \quad (3.1)$$

$$f_{\mathbf{X}|T}(\mathbf{X}=\mathbf{x}|T=1) = w^t(\mathbf{x}) \times f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=0,M=1) \quad (3.2)$$

In other words, after weighting, we want the distribution of the covariates among treated and comparison units in the matched sample ($M=1$), to be similar to the distribution of the covariates among the treated at the population level. From (3.2) we obtain that

$$\begin{aligned} w^t(\mathbf{x}) &= \frac{f_{\mathbf{X}|T}(\mathbf{X}=\mathbf{x}|T=1)}{f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=1,M=1)} \\ &= \frac{f_{M|T}(M=1|T)}{f_{M|(\mathbf{X},T)}(M=1|\mathbf{X}=\mathbf{x},T=1)} \end{aligned} \quad (3.3)$$

If we do not trim any treated units from the survey sample, it holds that $f_{M|T}(M=1|T=1) = f_{SR|T}(SR=1|T=1)$ and $f_{M|\mathbf{X},T}(M=1|\mathbf{X}=\mathbf{x},T=1) = f_{SR|\mathbf{X},T}(SR=1|\mathbf{X}=\mathbf{x},T=1)$. Thus (3.3) can be expressed as:

$$\omega^t(\mathbf{x}) = \frac{1}{f_{SR|\mathbf{X},T}(SR=1|\mathbf{X}=\mathbf{x},T=1)}. \quad (3.4)$$

Therefore we can conclude that units in the treatment group should be weighted using the survey weights assigned by the survey design. Combining (3.1), (3.2) and (3.4)

allows us to find an expression for the weights of the comparison units:

$$\begin{aligned} w^c(\mathbf{x}) &= \omega^t(\mathbf{x}) \times \frac{f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=1,M=1)}{f_{\mathbf{X}|(T,M)}(\mathbf{X}=\mathbf{x}|T=0,M=1)} \\ &= \omega^t(\mathbf{x}) \times \frac{f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)}{1-f_{(T|\mathbf{X},M)}(T=1|\mathbf{X}=\mathbf{x},M=1)} \times \frac{f_{(T|M)}(T=0|M=1)}{f_{(T|M)}(T=1|M=1)} \end{aligned} \tag{3.5}$$

where $f_{T|\mathbf{X},M}(T=1|\mathbf{X}=\mathbf{x},M=1)$ is the value of the propensity score computed among the matched observations. Since we implemented $k:1$ matching it holds that $\frac{f_{T|M}(T=0|M=1)}{f_{T|M}(T=1|M=1)} = \frac{\frac{k}{(k+1)}}{\frac{1}{(k+1)}} = k$, thus we can write (3.5) as

$$\omega^c(\mathbf{x}) = \omega^t(\mathbf{x}) \times \frac{f_{T|\mathbf{X},M}(T=1|\mathbf{X}=\mathbf{x},M=1)}{1-f_{T|\mathbf{X},M}(T=1|\mathbf{X}=\mathbf{x},M=1)} \times k$$

Also note that for a large matched sample, it should hold that

$$\frac{f_{T|\mathbf{X},M}(T=1|\mathbf{X}=\mathbf{x},M=1)}{1-f_{T|\mathbf{X},M}(T=1|\mathbf{X}=\mathbf{x},M=1)} = \frac{1}{k}$$

Thus $\omega^c(\mathbf{x}) = \omega^t(\mathbf{x})$.

This suggests that the matched comparison units should be assigned the survey weight of the treatment unit they have been matched to. Thus, the weights of the units in the comparison group are different from their original survey weights. Details of the resulting estimator of the PATT using this weight transfer, is available in the supplementary material (see Appendix C).

3.3 Simulation Study

In order to explore the empirical implications of the results of the previous section, we implement a simulation study to assess (1) whether the performance of the propensity score matching estimator is affected by how (or if) the survey weights are incorporated in the estimation of the propensity score model, (2) whether the weight transfer presented in Section 3.2.4 improves the performance of matching estimators, and (3) whether our conclusions depend on the assumed non-response mechanism and on the difference between the SATT and the PATT.

Our simulation set-up follows closely the one used by Austin et al. (2016). Austin et al. (2016) consider a population of size $N = 1,000,000$, with 10 strata. In each stratum there are 20 clusters, each composed of $5,000$ units. Six baseline covariates $(X_1, ..., X_6)$ are considered. The data generating mechanism for the baseline covariates is such that: (1) the probability density function is normal, (2) the covariates are independent (i.e., correlation between any pair of covariates is set equal to 0), (3) the standard deviation, across all the covariates, is equal to 1 and (4) the means vary across strata and cluster. More explicitly, for each strata (j), the mean of the covariates deviates in μ_{lj} from 0, where μ_{lj} are obtained assuming that $\mu_{lj} \sim N(0, \tau^{stratum})$. Within each strata, the mean of each cluster (k) deviates from the strata specific mean by μ_{lk}, with $\mu_{lk} \sim N\left(0, \tau^{cluster}\right)$. Thus the distribution of the l^{th} variable, in the j^{th} stratum, among the units of the k^{th} cluster is $X_{l,ijk} \sim N(\mu_{lj} + \mu_{lk},\ 1)$. We set $\tau^{stratum} = 0.35$ and $\tau^{cluster} = 0.25,\ 0.15,\ 0.05$. Each value of $\tau^{cluster}$ defines a

Scenario 1, 2 and 3 respectively. The probability of receiving treatment depends on the baseline covariates via a logistic model. The conditional means of the potential outcomes are constructed as a linear function of treatment, baseline covariates and interactions terms between T and X_1, X_2 and X_3 (i.e., the treatment effect is heterogeneous). In our simulation study, we introduce stratum-specific treatment effects, which allow us to vary the difference between the PATT and the SATT. The size of the stratum-specific effects are selected such that $\left(\frac{SATT}{PATT} - 1\right) \times 100$ takes roughly the values $-50\%, -40\%, -30\%, -20\%, -10\%$ and 0%.

We also extend the original set-up by considering four **non-response** scenarios. The first two scenarios are No-missing data (**NM**) and Missing at Random (**MAR**) where non-response depends on the six baseline covariates. The third is Missing at Random with additional covariate (**MARX**) where non-response depends on the same six baseline covariates plus an additional covariate X_7 not observed in the survey sample. In this situation, the survey weights are constructed using all seven covariates, but the analysis uses only six; this reflects the reality that some data (e.g., number of contact attempts) may be available to the team that conducted the survey but are typically not available to data users. The fourth mechanism is Missing at Random where non-response depends on the six baseline covariates and the treatment assignment (**MART**). Across the four mechanisms, the probability of response is generated using logistic models.

The final survey weights are defined as the number of individuals each person

represents times the inverse of the probability of responding. The average response rate across the MAR, MARX and MART models is close to 90%. We are aware that this response rate is high, nevertheless this allows us to compare the performance of the different PATT estimators without incurring in sample size adjustments. In practice, to compensate for non-response, samples sizes are increased by the inverse of the average response rate. By considering a relatively high respond rate, we do not need to implement such adjustments. We believe that increasing the non-response rate will only exacerbate our results.

For each scenario, and for each level of PATT-SATT relative difference, we ran 1000 iterations, and compare the performance of different estimators (described shortly). Performance is quantified by three metrics: (1) bias (in absolute value), (2) root mean square error (RMSE, defined as the square root of the sum of the squared bias and the variance of the estimator) and (3) empirical coverage of the 95% confidence interval. The parameters chosen in the simulation study are generally the same as the ones used by Austin et al. (2016) – see details in Appendix B in the supplementary material.

3.3.1 Estimators of the PATT

The estimators of the PATT considered in our article are grouped based on: (1) how the survey weights are used in the estimation of the propensity score and (2) whether the weight transfer described in Section 3.2.4 is implemented. Regarding the

estimation of the propensity score, there are three alternatives to consider regarding the use of the survey weights: (1) not incorporate the weights in the estimation (**UPS**), (2) incorporate the weights in a weighted estimation (**WPS**), and (3) incorporate the survey weights as a covariate in the propensity score model (**CPS**). Once the propensity score was estimated, 1:1 matching (nearest neighbor) without replacement was implemented. After the matching procedure is executed, the survey weight of the treated can be transferred to the comparison units they have been matched to (**WT**) or each observation can retain their original survey weights (**OW**). Therefore, the estimator labeled as "**CPS|WT**" is the estimator of the PATT in which the survey weights are used as a covariate in the estimation of the propensity score model and the weight transfer described in Section 3.2.4 is implemented.

In addition to the 6 estimators previously described, we also considered a "**Naïve**" estimator of the PATT. The Naïve estimator uses propensity score matching but ignores the survey design all together. That is, it does not incorporate the survey weights in the estimation of the propensity score nor uses them to weight the observed outcome; the Naïve estimator is thus a valid estimator of the SATT but not necessarily the PATT.

Work by Cochran and Rubin (1973), Rubin (1973b), Carpenter (1977), Rubin (1979), Rosenbaum and Rubin (1984), Robins and Rotnitzky (1995), Heckman and Todd (2009), Rubin and Thomas (2000), Glazerman et al. (2003) Imai and Van Dyk (2004), Abadie and Imbens (2006) and Ho et al. (2007) suggested that defining an

outcome model that adjusts for confounders in the estimation of causal effects, can improve causal inferences. Thus, following Ridgeway et al. (2015) we considered two outcome models: (1) an "unadjusted" model that has the treatment assignment (T) as the only regressor and (2) an "adjusted" model that in addition to the treatment assignment, includes the baselines covariates as regressors, but it does not include interaction terms.

3.3.2 Results

Diagnostics

First we evaluate how balanced the distribution of the survey weights and the baseline covariates is between the treated and comparison groups as a result of implementing the matching procedures described in Section 3.3.1. Balance is defined in terms of the standardized mean difference (SMD). Notice that the SMD under the Naive estimation approach provides a measure of balance achieved in the sample, since it does not incorporate the survey weights. Since the other methods do incorporate survey weights in their estimation of causal effects, the calculation of the SMD associated with these estimators uses the survey weights, therefore we consider them measures of balance at the population level. Table 3.1, shows the SMD in the population, before any matching procedure was implemented.

Figures 3.1, 3.2 and 3.3 summarize our main findings for Scenarios 1, 2 and 3,

respectively. In these figures, the vertical axis displays the average value of the SMD (across the $1,000$ iterations in our simulation study). Each row of plots represents a different non-response scenario. Each shape identifies the different procedure used in the estimation of the propensity score model: (1) triangles are associated with the estimators that do not incorporate the survey weights in the estimation of the propensity score model, but the sample weights are used in the computation of the SMD after matching, (2) circles represents the SMD after matching when the survey weights were incorporated in a weighted estimation of the propensity score model and (3) squares show the balance achieved after the matching procedure when the weights were used as a covariate in the estimation of the propensity score model. Darker shaded markers are associated with the implementation of the weight transfer described in Section 3.2.4, lighter shades show the balance achieved when each sample unit kept its original survey weight. We also display the SMD achieved by the Naive estimator using a black asterisk. The red line in Figure 3.1, shows the threshold value of 0.20 (see Rosenbaum and Rubin (1985)); SMDs above that threshold indicate that the matching procedure was not effective.

The patterns that we observe in Figure 3.1 are consistent across all scenarios (see Figures 3.2 and 3.3). First we observe that, in general, good balance is achieved by all matching procedures, although there are some exceptions; the SMD for covariate X_6 is not always below 0.20 when the non-response mechanism is MART. Second, when the non-response mechanism is MAR, MARX or No Missing, the weight transfer may

translate into worse balance (this is particularly clear for covariate X_3 and in multiple covariates in Scenario 3). Nevertheless, this situation is reversed when the non-response mechanism is MART. In fact, failure to implement the weight transfer can yield poor balance in some of the covariates. Interestingly we observe that when the non-response mechanism is different from MART, balance in the covariates translates into balance of the survey weights.

Finally note that for most of the baseline covariates and across non-response mechanisms the Naive method achieves better balance than any other of the matching procedures implemented. However it is important to notice that the Naive method achieves good balance in the sample, but this does not imply that good balance is achieved in the population.

Treatment Effect Estimation Results

The estimators that only used the treatment indicator (T) as a covariate in the outcome model estimation are labeled as "$\mathbf{Y} \sim \mathbf{T}$", whereas the estimators that additionally adjust for the vector of covariates $\mathbf{X}$ are labeled as "$\mathbf{Y} \sim \mathbf{T} + \mathbf{X}$"

Figure 3.4 displays our findings for Scenario 1. Each column in the plot shows one of the metrics (bias in absolute value, empirical coverage of the 95% confidence interval and RMSE) chosen to asses the performance of the different matching estimators. Each row of plots represent a different non-response scenario. We keep the same shape scheme used to assess the balance. The type of line is associated with how

the survey weights we incorporated in the outcome estimation. Solid, darker lines are associated with the implementation of the weight transfer, whereas dashed lighter lines show the performance achieved when each sample unit kept its original sampling weight. The performance achieved by the Naive estimator is shown using solid black lines with asterisks.

Differences in the performance of the estimators are more pronounced when we consider the "unadjusted" estimators. As expected, as the percentage difference between the SATT and the PATT increases in absolute value, the naive estimator performance worsens. Notice that this result holds even when the outcome model adjusts for the covariates. When survey weights are incorporated in the analysis keeping the original weights translates to reduction of bias (this is true for all non-response mechanisms considered except MART). Adjusting for covariates in the outcome model translates into better performance (across the three metrics considered). In general we observe that how the survey weights are incorporated in the estimation of the propensity score does not yield differences in the performance of the estimators. When the non-response model is MART we observe that the weight transfer reduces bias associated with the estimation of the PATT; this is true even after adjusting for relevant covariates (although is more obvious among the "unadjusted" estimators of the PATT). Furthermore, among the "unadjusted" estimators, we observe that the weight transfer is not only associated with better balance but also better coverage and better RMSE. Among the "unadjusted" estimators and when the non-response

mechanisms is MART, we also observe that a weighted estimation of the propensity score models translates into gains of efficiency (reduction of the RMSE). Nevertheless, this gain is not substantial when covariates are included in the outcome model. We believe that the reason why the weight transfer described in Section 3.2.4 does not improve the performance of the estimators in the non-response mechanisms besides MART (i.e., MAR, MARX and No-Missing) is due to the fact that if the matching procedure is successful, then balance in the covariates will translate in balance of the survey weights. Therefore the weight transfer is implicitly implemented. This hypothesis seems to be confirmed by Figures 3.1, 3.2 and 3.3, which show that when that non-response mechanism is different from MART, balance in the covariates translates into balance of the survey weights. Furthermore, notice that when the non-response is MART good balance of the baseline covariates does not imply good balance of the survey weights, and therefore the weight transfer improves the performance of the estimators. Another key feature of the results depicted in Figure 3.4 (and also true in Figures 3.5 and 3.6), is that even when the percentage difference between the SATT and the PATT is as high as 50%, incorporating the survey weights translates into significant bias reduction. However, as the percentage difference between the SATT and the PATT gets close to 0, no significant differences in the performance of the naive and the other estimators is observed (this is the default scenario of the simulation set-up implemented by Austin et al. (2016)).

3.4 Application

In this section we use The Early Childhood Longitudinal Study, Kindergarten class 1998-1999 (ECLS-K) (see Tourangeau et al. (2009)) to estimate the effect of special education services on math skills, replicating Keller and Tipton (2016). Keller and Tipton (2016) provide an excellent guide on how to implement different R packages to estimate causal effects by implementing different matching procedures using the work of Morgan et al. (2008) as a motivating example. We follow closely the work by Keller and Tipton (2016) since they provide a comprehensive list of the variables used in their analysis. It is worth noticing that Morgan et al. (2008) does not explicitly mention how the survey weights are incorporated in the estimation of the propensity score matching estimators and neither does Keller and Tipton (2016) (although Keller and Tipton (2016) explicitly state that the purpose of their article is to illustrate how different software can be use to implement propensity score matching and their results should not be interpreted in a causal context)

The ECLS-K is a longitudinal study that examines child early school experiences beginning with kindergarten until eighth grade that collects information: (1) at the child level , (2) at the household level and (3) at the school level. The data was accessed through `http://www.researchconnections.org/childcare/studies/28023` (see U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. (2011))

Since our goal is methodological, to compare the different methods, we do not

assess the plausibility of the key assumptions that would be needed to interpret the results as causal, and thus the results should not be treated as definitive causal effects regarding the effect of special of education services on learning skills. We follow Keller and Tipton (2016) and estimate the effect of elementary school special education services on math achievement in fifth grade. We consider 39 covariates in the propensity score model (which include demographic, socio-economical, academic, household and school level variables). A codebook of the variables used in this application is available in the supplementary materials (see SUP_Application.R). We fit an unadjusted outcome model as well as one that adjusts for the same set of covariates included in the propensity score model. Table **??** displays the balance achieved by each method considered in our simulation study. Overall we observe that most of the matching procedures were effective in increasing the balance for the covariates. Nevertheless, some of the methods were not able to improve balance enough to generate SMDs smaller than 0.20 on some of the covariates (see the highlighted cells in Table **??**). Notice that **WPS|WT** is the only method that achieved SMDs smaller that 0.20 in 38 of the 39 covariates considered. The last row in Table **??** shows the SMD of the survey weights after the matching procedure. Note that the good balance in the covariates does translate into good balance in the survey weights (across all methods); this seems to indicate that the non-response mechanism may not be MART and thus that the weight transfer may not improve estimation of the PATT.

Table 3.3, shows the estimated PATT. As expected most of the estimators produce

similar estimators with the exception of the Naive estimator, which is likely a biased estimate of the PATT.

3.5 Discussion

In this article we explore how different ways of using survey weights can affect the performance of propensity score matching PATT estimators based on complex survey data when different non-response mechanisms are considered. To our knowledge, this is the first article that explores the impact of non-response mechanisms on the performance of propensity score matching estimators.

We have also evaluated how the difference between that SATT and the PATT affect the performance of different propensity score matching estimators. When we first replicated the simulation study designed by Austin et al. (2016) we found that the Naïve estimator of the PATT performed as well as any of the other PATT estimators considered by the authors. This was due to the fact that the PATT and the SATT where practically identical. Based on our simulation study and application to the ECLS-K dataset, we conclude that:

How the survey weights are incorporated in the estimation of the propensity score does not affect the performance of the matching estimators. This result holds true across all non-response mechanisms, although we found evidence that a weighted estimation of the propensity score model can increase the efficiency of the

PATT estimator when an unadjusted outcome model is estimated and the missing data pattern is MART.

Adjusting for relevant covariates in the outcome model improves the performance of the estimators. This result is consistent with findings by others (e.g., Drake, 1993; McCaffrey et al., 2004; Frölich, 2007; Robins et al., 2007; Lee et al., 2011; Imai and Ratkovic, 2014; Ridgeway et al., 2015).

Survey weights should be incorporated in the outcome analysis. Our results indicate that not including survey weights in the estimation procedure may lead to substantial bias.

A weight transfer improves the performance of the matching estimators under the MART non-response mechanism. This performance improvement occurs when the PATT is estimated using an unadjusted outcome model.

Population balance of covariates is crucial to the estimation of population treatment effects. We found that the key element to obtain accurate estimates of the PATT is to achieve good **population balance** in the observed covariates. That is, survey weights need to be incorporated when assessing balance. Population balance (evaluated by SMD) was the best predictor of the performance of the estimator. In our simulation study we observe that the average correlation (i.e., averaged across the covariates) between bias and SMD achieved by the estimators that use survey weights (i.e., excluding the Naive estimator) is 0.77; the correlations (also excluding the Naive estimator) of Coverage and RMSE with SMD are -0.66

and 0.62, respectively. For the Naive estimator, we also computed the correlation of SMD (in this case, representing sample balance) with the three performance metrics; the correlations are 0.00 with Bias, -0.15 with Coverage, and -0.1 with RMSE. This shows that good sample balance does not necessarily translate into good performance of the propensity score matching estimator; it is population balance that matters.

The balance achieved in the survey weights after the matching procedure could potentially help identify the nature of the non-response mechanism. When the non-response is MART we observed that: (1) good balance in the confounders does not imply balance of the survey weights, and (2) the weight transfer improves the performance of the estimators. We therefore recommend checking **population balance** on the covariates and on the survey weights after matching. If balance is achieved on the former but not on the latter, and especially if there is theoretical or prior empirical basis to suspect that non-response (or sample selection) may have been influenced by treatment status, we recommend implementing a weight transfer.

It is important to note that the confidence intervals presented in this article were constructed using the 'survey' package in R (**?**). There has been limited work to evaluate the asymptotic properties of matching estimators, except for the significant contributions made by Abadie and Imbens (2006), Abadie and Imbens (2008) and Abadie and Imbens (2016). Future work will focus on generalizing their results to the context of complex survey data.

In our article we have restricted our attention to matching estimators of the PATT where the matching procedure was implemented without replacement. It has been pointed out – in the context of a SRS – that when matching with replacement is used, weights should created to guarantee that the matched treated and comparison groups are weighted up to be similar (Ho et al., 2011); future work will extend such weights computation to cover matching with replacement using complex survey data. Finally, in our simulation study we assume that the propensity score model is correctly specified. Future work will evaluate how the performance of propensity score matching estimators is affected by misspecification of the propensity score model, of the outcome model, and of both.

In conclusion, accurate estimates of the PATT can be obtained using complex survey data and propensity score matching, especially if it can be shown that good covariate balance is obtained in the population of interest.

Acknowledgments

The authors wish to thank Francis M. Abreu, whose comments greatly improved this manuscript. This work was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305D150001 (PIs: Stuart and Dong). *Conflict of Interest*: None declared.

Software

The simulation study, plots and the application were implemented using the software R and the platform RStudio (RStudio Team, 2015). The following packages were used: 'data.table' (Dowle et al., 2015), 'ggplot2' (Wickham, 2009), 'MatchIt' (Ho et al., 2011), 'survey' (**?**), 'sampling' (Tillé and Matei, 2015) and 'xtable' (Dahl, 2016).

Supplementary Material

Supplementary material include: Appendices A, B, C and D. The R script used to: (1) generate the population data (SUP_DataGen.R), (2) implement the simulation study (SUP_Simulation.R), (3) combine the results (SUP_CombiningResults.R), (4) create plots (SUP_PlotSMD.R and SUP_PlotPATT.R) and (5) execute the application (SUP_Application.R this R script also contains the codebook for the variables used) and is available online at `http://biostatistics.oxfordjournals.org`.

Bibliography

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Akee, R. (2011). Errors in self-reported earnings: The role of previous earnings volatility and individual characteristics. *Journal of Development Economics*, 96(2):409–421.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

BIBLIOGRAPHY

Austin, P. C., Jembere, N., and Chiu, M. (2016). Propensity score matching and complex surveys. *Statistical Methods in Medical Research*, page 0962280216658920.

Bound, J. and Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, pages 1–24.

Brunell, T. L. and DiNardo, J. (2004). A propensity score reweighting approach to estimating the partisan effects of full turnout in american presidential elections. *Political Analysis*, 12(1):28–45.

Carpenter, R. (1977). Matching when covariables are normally distributed. *Biometrika*, pages 299–307.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the simex estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433):242–250.

Cochran, W. G. (1977). Sampling techniques. 1977. *New York: John Wiley and Sons*.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International journal of epidemiology*, 35(4):1074–1081.

BIBLIOGRAPHY

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, 89(428):1314–1328.

Dahl, D. B. (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2.

Dowle, M., Srinivasan, A., Short, T., with contributions from R Saporta, S. L., and Antonyan, E. (2015). *data.table: Extension of Data.frame*. R package version 1.9.6.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, pages 1231–1236.

Edwards, J. K., Cole, S. R., Westreich, D., Crane, H., Eron, J. J., Mathews, W. C., Moore, R., Boswell, S. L., Lesko, C. R., Mugavero, M. J., et al. (2015). Multiple imputation to account for measurement error in marginal structural models. *Epidemiology*, 26(5):645–652.

Frölich, M. (2007). Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the united kingdom. *The Econometrics Journal*, 10(2):359–407.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, pages 153–164.

BIBLIOGRAPHY

Glazerman, S., Levy, D. M., and Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, 589(1):63–93.

Goetghebeur, E. and Vansteelandt, S. (2005). Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research*, 14(4):397–415.

Goodman, E. and Whitaker, R. C. (2002). A prospective study of the role of depression in the development and persistence of adolescent obesity. *Pediatrics*, 110(3):497–504.

Grace, Y. Y. (2008). A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*, 9(3):501–512.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey methodology*, volume 561. John Wiley & Sons.

Guo, Y., Little, R. J., and McConnell, D. S. (2012). On using summary statistics from an external calibration sample to correct for covariate measurement error. *Epidemiology*, 23(1):165–174.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

BIBLIOGRAPHY

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793.

Harder, V. S., Stuart, E. A., and Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3):234.

Harris, K. M., Halpern, C., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and J.R., U. (2009). The national longitudinal study of adolescent to adult health: Resear design. *Availabe at: http://www.cpc.unc.edu/projects/addhealth/design Accesed May 15, 2015.*

Heckman, J. J. and Todd, P. E. (2009). A note on adapting propensity score matching and selection models to choice based samples. *The econometrics journal*, 12(s1):S230–S234.

Heid, I., Küchenhoff, H., Miles, J., Kreienbrock, L., and Wichmann, H. (2004). Two dimensions of measurement error: classical and berkson error in residential radon exposure assessment. *Journal of Exposure Science and Environmental Epidemiology*, 14(5):365–377.

Hernan, M. A. and Robins, J. M. (2017). *Causal Inference*. Boca Raton: Chapman & Hall/CRC. Forthcoming.

BIBLIOGRAPHY

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4):272–279.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539.

BIBLIOGRAPHY

Keller, B. and Tipton, E. (2016). Propensity score analysis in ra software review. *Journal of Educational and Behavioral Statistics*, page 1076998616631744.

Korn, E. L. and Graubard, B. I. (1995a). Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 263–295.

Korn, E. L. and Graubard, B. I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49(3):291–295.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174.

Lenis, D., Ebnesajjad, C. F., and Stuart, E. A. (2017a). A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error. *Biostatistics*, 18(2):325–337.

Lenis, D., Nguyen, T. Q., Dong, N., and Stuart, E. A. (2017b). It's all about balance: Propensity score matching in the context of complex survey data. *Under Review*.

Little, R. (2003). The bayesian approach to sample survey inference. *Analysis of Survey Data*, pages 49–57.

Little, R. J. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326.

BIBLIOGRAPHY

Lockwood, J. and McCaffrey, D. (2015). Simulation-extrapolation for estimating means and causal effects with mismeasured covariates. *Observational Studies*, 1:241–290.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19. R package verson 2.2.

Lumley, T. (2016). survey: analysis of complex survey samples. R package version 3.32.

McCaffrey, D. F., Lockwood, J., and Setodji, C. M. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, page ast022.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

Morgan, P. L., Frisco, M. L., Farkas, G., and Hibel, J. (2008). A propensity score matching analysis of the effects of special education services. *The Journal of special education*.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Stat Sci*, 5:463–472.

of Education. Institute of Education Sciences. National Center for Education Statistics., U. S. D. (2011). Early childhood longitudinal study [united states]: Kindergarten class of 1998-1999, kindergarten-eighth grade full sample. icpsr28023-v1. ann arbor, mi: Inter-university consortium for political and social research [distributor]. http://doi.org/10.3886/ICPSR28023.v1http://doi.org/10.3886/ICPSR28023.v1.

Pettersen, B. J., Anousheh, R., Fan, J., Jaceldo-Siegl, K., and Fraser, G. E. (2012). Vegetarian diets and blood pressure among white subjects: results from the adventist health study-2 (ahs-2). *Public health nutrition*, 15(10):1909–1916.

Plankey, M. W., Stevens, J., Fiegal, K. M., and Rust, P. F. (1997). Prediction equations do not eliminate systematic error in self-reported body mass index. *Obesity research*, 5(4):308–314.

Reardon, S. F., Cheadle, J. E., and Robinson, J. P. (2009). The effect of catholic schooling on math and reading development in kindergarten through fifth grade. *Journal of Research on Educational Effectiveness*, 2(1):45–87.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., and Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, 3(2):237–249.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Perfor-

mance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559.

Robins, J. M. (2003). General methodological considerations. *Journal of Econometrics*, 112(1):89–106.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rosner, B., Spiegelman, D., and Willett, W. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case

of multiple covariates measured with error. *American Journal of Epidemiology*, 132(4):734–745.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

BIBLIOGRAPHY

Rubin, D. B. and Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics*, pages 1814–1826.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.

Saint-Maurice, P. F., Welk, G. J., Beyler, N. K., Bartee, R. T., and Heelan, K. A. (2014). Calibration of self-report tools for physical activity research: the physical activity questionnaire (paq). *BMC public health*, 14(1):1.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Steiner, P. M., Cook, T. D., and Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2):213–236.

Stommel, M. and Schoenborn, C. A. (2009). Accuracy and usefulness of bmi measures

based on self-reported weight and height: findings from the nhanes & nhis 2001-2006. *BMC public health*, 9(1):1.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology*, 162(3):279–289.

Tillé, Y. and Matei, A. (2015). *sampling: Survey Sampling*. R package version 2.7.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., and Najarian, M. (2009). Early childhood longitudinal study, kindergarten class of 1998-99 (ecls-k): Combined user's manual for the ecls-k eighth-grade and k-8 full sample data files and electronic codebooks. nces 2009-004. *National Center for Education Statistics*.

Webb-Vargas, Y., Rudolph, K. E., Lenis, D., Murakami, P., and Stuart, E. A. (2015). Applying multiple imputation for external calibration to propensty score analysis.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wingood, G. M., DiClemente, R. J., Harrington, K., and Davies, S. L. (2002). Body

image and african american females' sexual health. *Journal of women's health & gender-based medicine*, 11(5):433–439.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of data Science*, 4(1):67–91.

Table 3.1: *Standardized Mean Differences (Population level)*

Scenario	X1	X2	X3	X4	X5	X6
1	0.11	0.28	0.43	0.49	0.69	0.81
2	0.03	0.16	0.34	0.59	0.57	0.91
3	0.09	0.22	0.33	0.48	0.60	0.81

Table 3.2: SMD achieved by the different estimation procedures.

Variable	Naive	UPS		CPS		WPS	
		OW	WT	OW	WT	OW	WT
FEMALE	0.08	0.08	0.06	0.08	0.06	0.03	0.01
WHITE	0.05	0.05	0.05	0.05	0.05	0.10	0.14
WKSESL	0.04	0.04	0.06	0.04	0.06	0.05	0.09
C1R4RSCL	0.04	0.04	0.03	0.04	0.03	0.06	0.04
C1R4MSCL	0.13	0.13	0.24	0.13	0.24	0.00	0.04
S2KPUPRI	0.25	0.25	0.15	0.25	0.15	0.01	0.02
P1ELHS	0.02	0.02	0.04	0.02	0.04	0.02	0.10
P1EHS	0.06	0.06	0.09	0.06	0.09	0.05	0.03
P1ESC	0.10	0.10	0.08	0.10	0.08	0.02	0.00
P1EC	0.15	0.15	0.09	0.15	0.09	0.13	0.04
P1EMS	0.00	0.00	0.07	0.00	0.07	0.08	0.04
P1EPHD	0.04	0.04	0.00	0.04	0.00	0.10	0.05
P1FIRKDG	0.16	0.16	0.14	0.16	0.14	0.20	0.17

Continued on next page

Table 3.2 – continued from previous page

Variable	Naive	UPS		CPS		WPS	
		OW	WT	OW	WT	OW	WT
P1AGEENT	0.12	0.12	0.07	0.12	0.07	0.06	0.06
T1LEARN	0.01	0.01	0.05	0.01	0.05	0.05	0.10
P1HSEVER	0.03	0.03	0.03	0.03	0.03	0.03	0.16
FKCHGSCH	0.00	0.00	0.09	0.00	0.09	0.05	0.12
S2KMINOR	0.10	0.10	0.09	0.10	0.09	0.21	0.12
P1FSTAMP	0.02	0.02	0.13	0.02	0.13	0.05	0.14
SGLPAR	0.05	0.05	0.07	0.05	0.07	0.04	0.17
TWOPAR	0.05	0.05	0.07	0.05	0.07	0.04	0.17
P1NUMSIB	0.06	0.06	0.01	0.06	0.01	0.07	0.12
P1HMAFB	0.04	0.04	0.17	0.04	0.17	0.03	0.19
WKCAREPK	0.03	0.03	0.14	0.03	0.14	0.06	0.06
P1EARLY	0.07	0.07	0.09	0.07	0.09	0.05	0.09
P1WEIGHO	0.06	0.06	0.11	0.06	0.11	0.05	0.09
C1FMOTOR	0.14	0.14	0.29	0.14	0.29	0.13	0.11
C1GMOTOR	0.15	0.15	0.20	0.15	0.20	0.06	0.07
P1HSCALE	0.12	0.12	0.08	0.12	0.08	0.04	0.05
P1SADLON	0.04	0.04	0.22	0.04	0.22	0.02	0.01

Continued on next page

Table 3.2 – continued from previous page

Variable	Naive	UPS		CPS		WPS	
		OW	WT	OW	WT	OW	WT
P1IMPULS	0.09	0.09	0.17	0.09	0.17	0.02	0.06
P1ATTENI	0.14	0.14	0.23	0.14	0.23	0.10	0.04
P1SOLVE	0.26	0.26	0.38	0.26	0.38	0.20	0.14
P1PRONOU	0.03	0.03	0.10	0.03	0.10	0.28	0.26
P1DISABL	0.13	0.13	0.08	0.13	0.08	0.12	0.04
AVG4RSCL	0.03	0.03	0.04	0.03	0.04	0.15	0.03
AVG4MSCL	0.01	0.01	0.04	0.01	0.04	0.19	0.02
AVGWKSES	0.03	0.03	0.06	0.03	0.06	0.14	0.03
C1_6FC0	0.11	0.11	0.00	0.11	0.00	0.08	0.00

Table 3.3: PATT estimation. Unadjusted vs. Adjusted

	Unadjusted	95% CI	Adjusted	95% CI
Naive	-2.62	(-4.44; -0.81)	-3.30	(-5.98; -0.61)
UPS|OW	-5.25	(-8.55; -1.94)	-7.86	(-13.42; -2.30)
UPS|WT	-4.33	(-7.24; -1.42)	-9.92	(-14.98; -4.86)
CPS|OW	-5.79	(-8.98; -2.61)	-6.63	(-12.18; -1.08)
CPS|WT	-5.31	(-8.39; -2.24)	-7.59	(-12.89; -2.29)
WPS|OW	-4.62	(-8.05; -1.19)	-6.39	(-11.90; -0.88)
WPS|WT	-2.80	(-6.13; 0.53)	-5.97	(-11.34; -0.61)

The first column displays the estimation result of implementing an unadjusted regression model and the second column shows the associated 95% confidence interval. The third column, shows the results of estimating the PATT adjusting for the set of covariates considered in Table **??**, and the last column shows the associated 95% confidence interval.

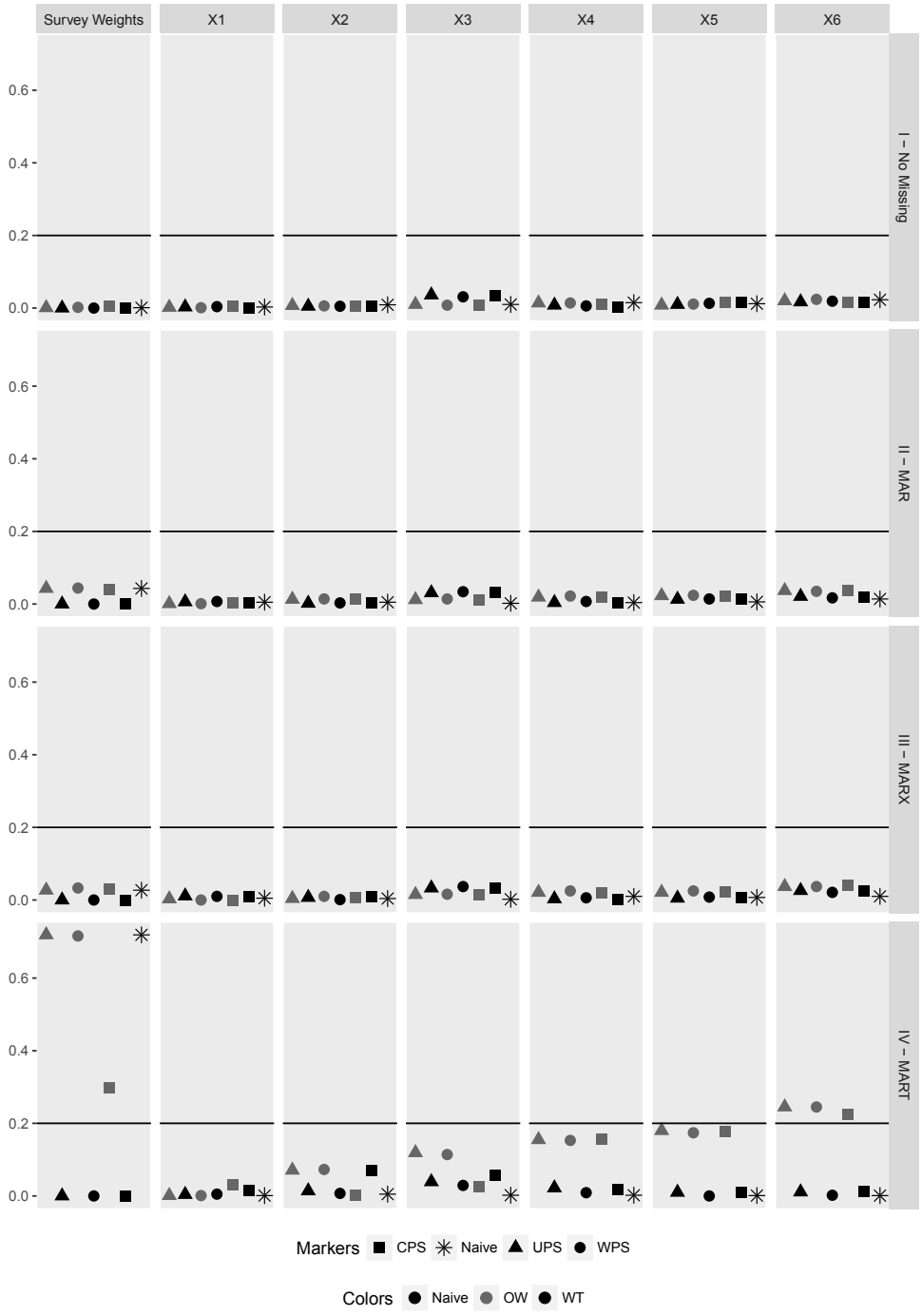

Figure 3.1: Diagnostics. SMD computed in the matched samples in Scenario 1. The Naive method represents balance in the sample (since it does not involve the survey weights in its computation). The other methods incorporate survey weights in the computation of the SMD, thus are considered measures of balance in the population.

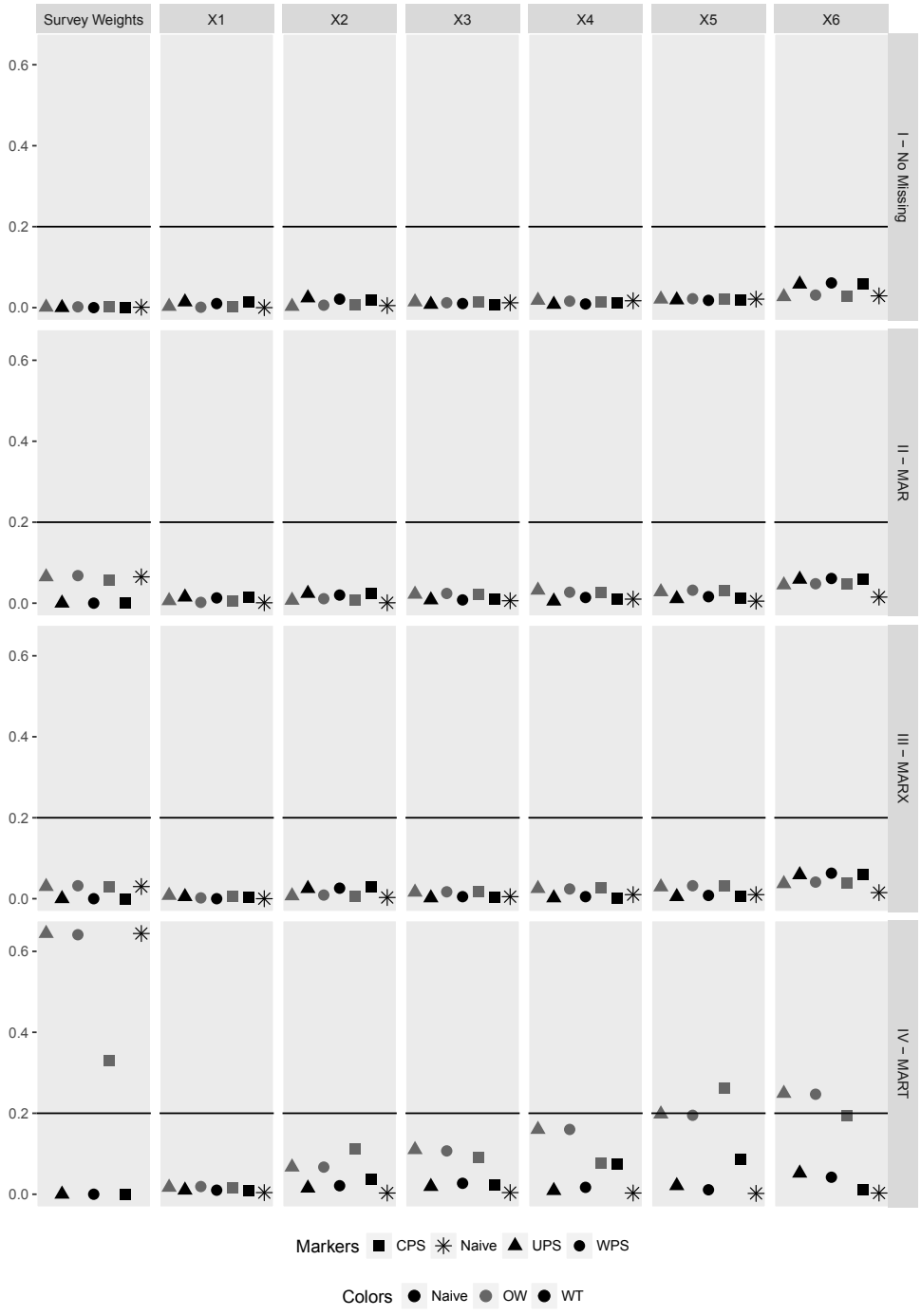

Figure 3.2: Diagnostics. SMD computed in the matched samples in Scenario 2. The Naive method represents balance in the sample (since it does not involve the survey weights in its computation). The other methods incorporate survey weights in the computation of the SMD, thus are considered measures of balance in the population.

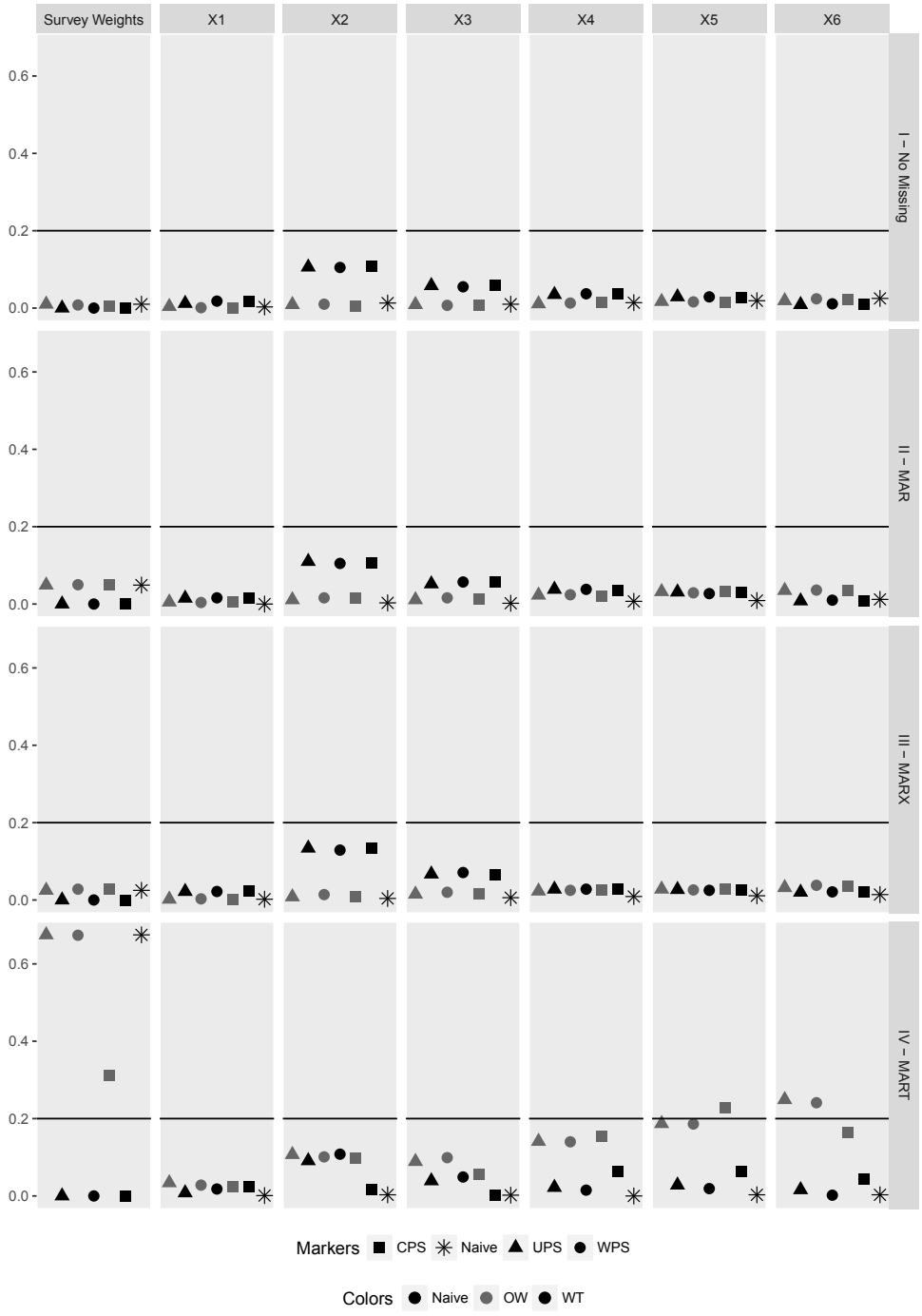

Figure 3.3: Diagnostics. SMD computed in the matched samples in Scenario 3. The Naive method represents balance in the sample (since it does not involve the survey weights in its computation). The other methods incorporate survey weights in the computation of the SMD, thus are considered measures of balance in the population.

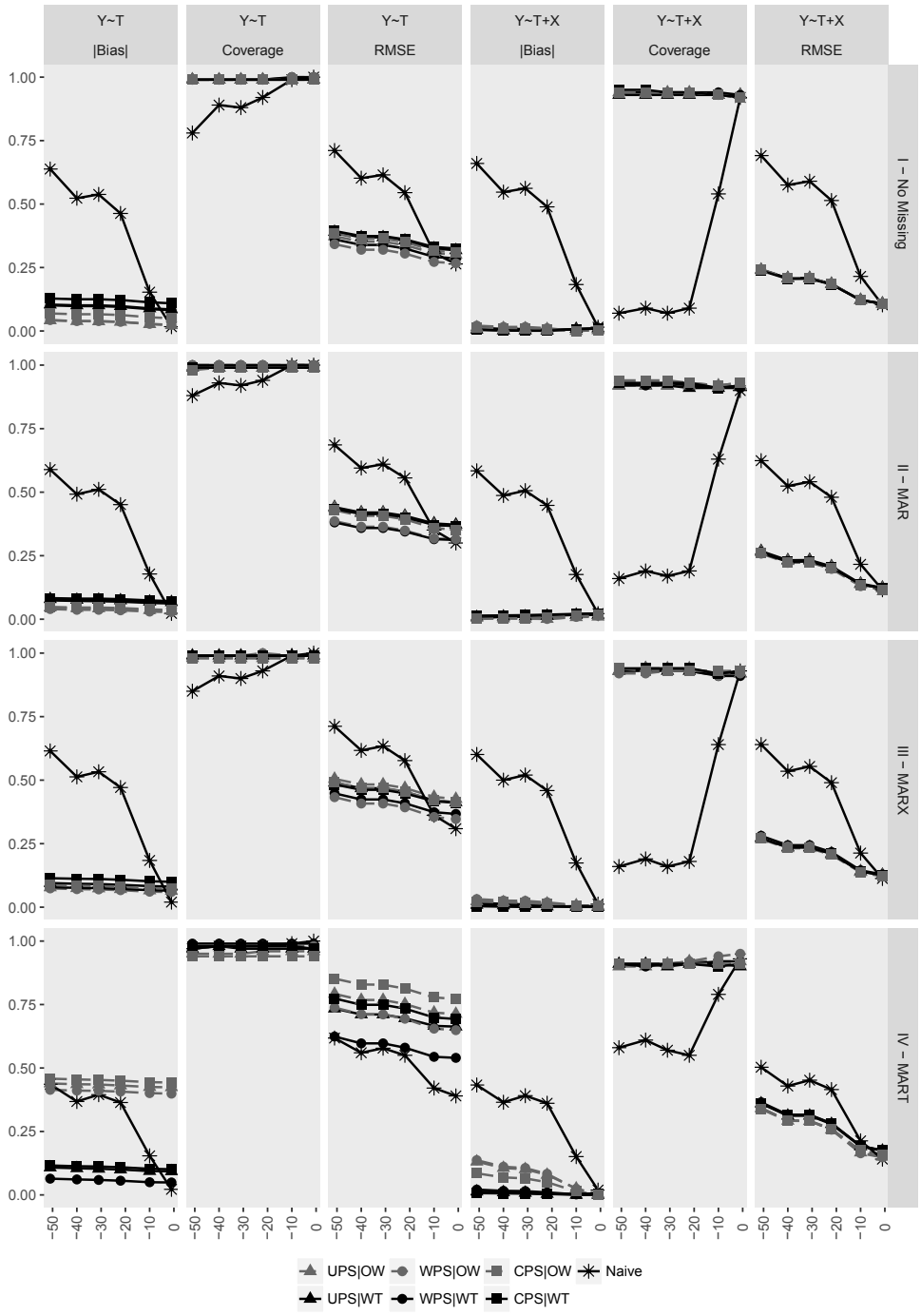

Figure 3.4: Scenario 1 Bias in absolute value, coverage and root mean squared error (RMSE) as functions of the % difference between the SATT and PATT (simulation study).

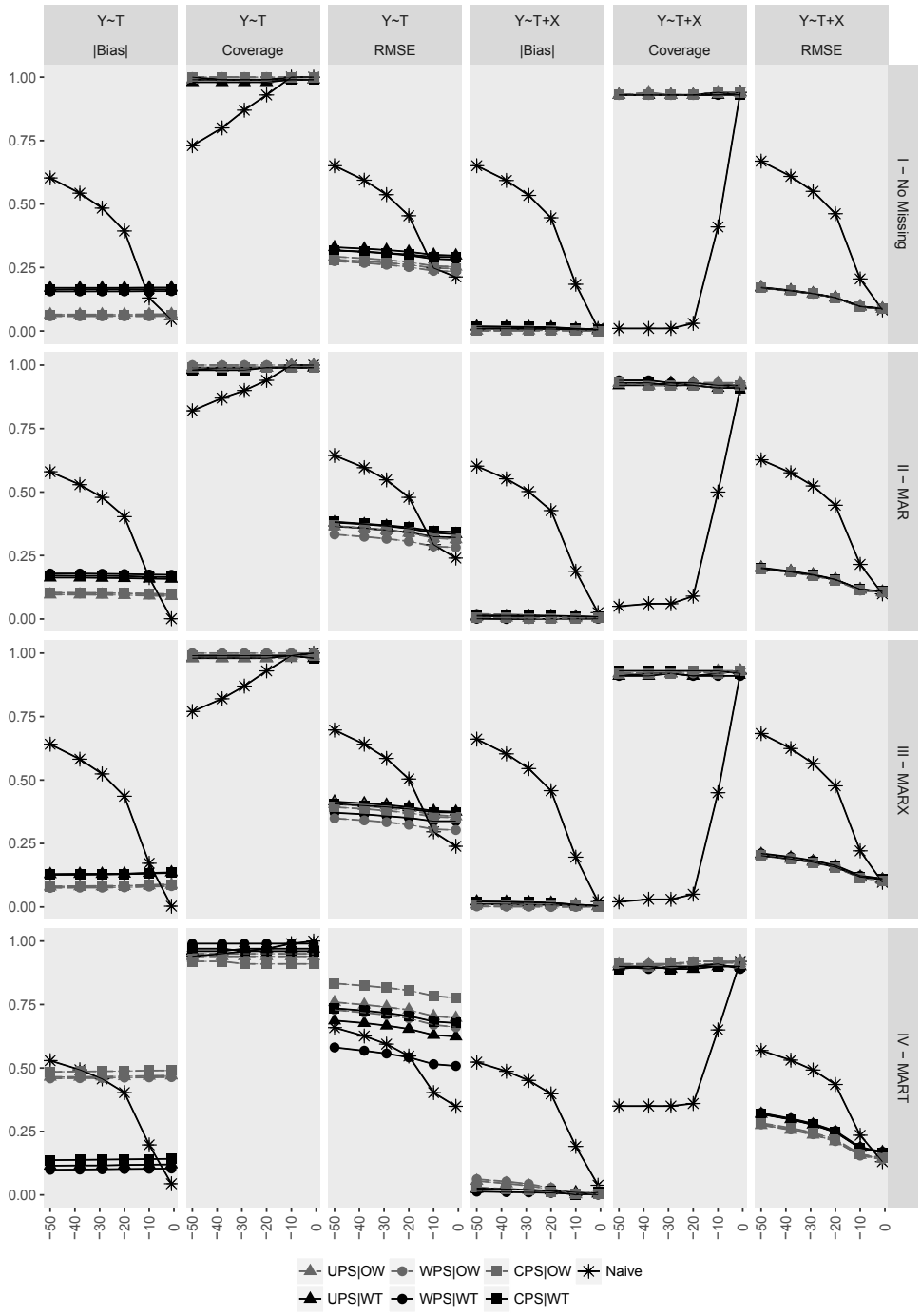

Figure 3.5: Scenario 2 Bias in absolute value, coverage and root mean squared error (RMSE) as functions of the % difference between the SATT and PATT (simulation study).

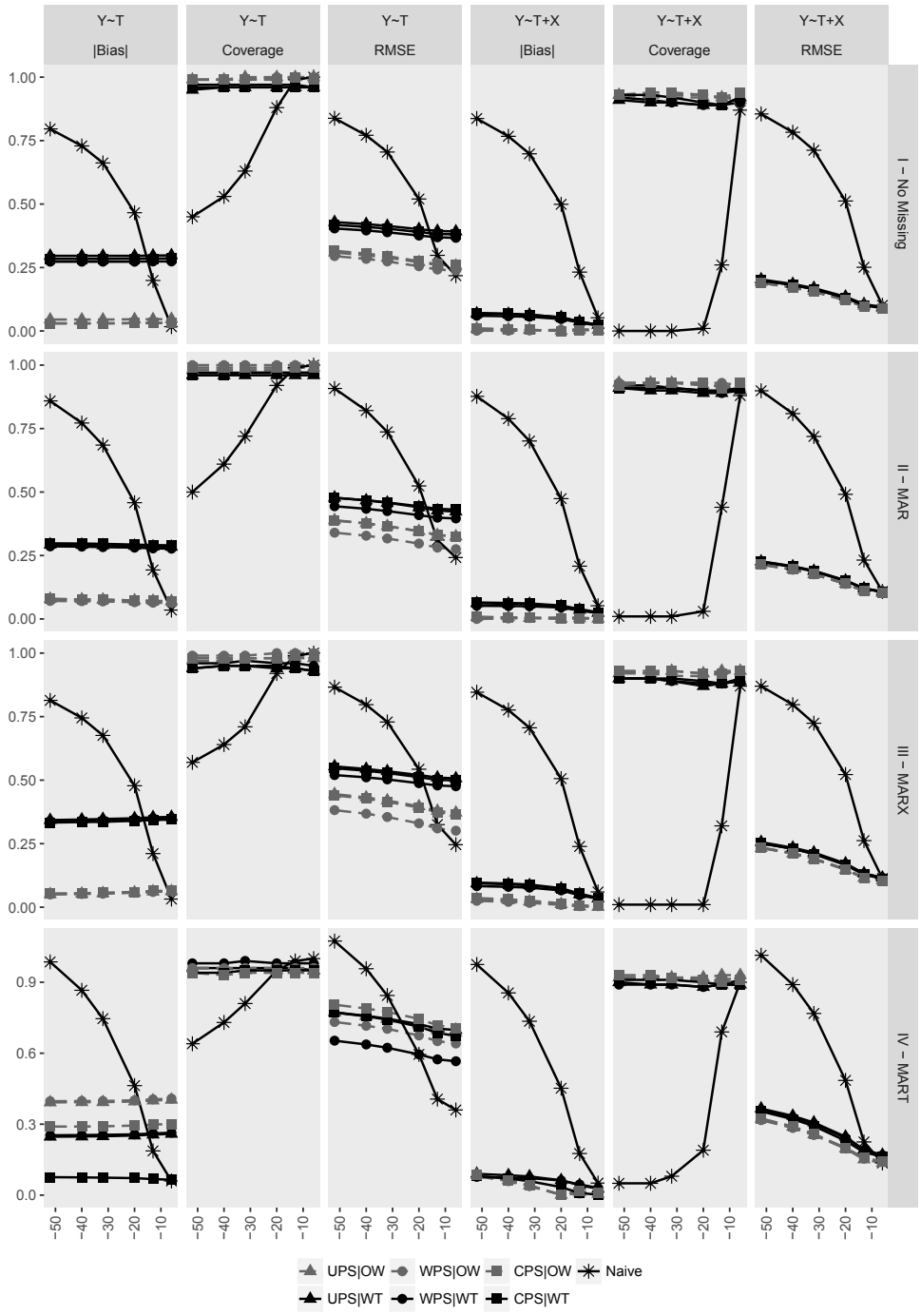

Figure 3.6: Scenario 3 Bias in absolute value, coverage and root mean squared error (RMSE) as functions of the % difference between the SATT and PATT (simulation study).

Chapter 4

Propensity Score Methods Under Different Degree of Model Misspecification in the Context of Complex Survey Data

4.1 Introduction

Randomized clinical trials (RCT) are considered the gold standard for estimating causal effects. In a RCT the researcher knows the treatment assignment mechanisms, allowing unbiased estimators of causal effects. Nevertheless, is not unusual to find circumstances where a random assignment of the treatment is unfeasible or unethical.

When this happens, researchers need to rely on non-experimental data.

A main drawback of non-experimental data is that the treatment assignment is not random, therefore there may be confounders that are related to the outcome and differ between treatment and comparison groups. Failure to address confounding will lead to biased estimators of causal effects. One way to mitigate confounding by observed characteristics is using the propensity score, which models the probability of being assigned to the treatment group given the set of confounders.

Model misspecification can be an issue in two ways when using propensity score methods to estimate causal effects: first, in estimating the propensity score, and second, in the outcome model. Since the true treatment assignment mechanism is hardly ever known when working with non-experimental data, different approaches have been suggested to model and estimate the propensity score. While some authors have proposed nonparametric estimation procedures (Hahn, 1998; Imbens, 2004; Ho et al., 2011), it is common practice to estimate the propensity score parametrically via logistic regression.

Models are also used in the outcome analysis. Work by Cochran and Rubin (1973), Rubin (1973b), Carpenter (1977), Rubin (1979), Rosenbaum and Rubin (1984), Robins and Rotnitzky (1995), Heckman and Todd (2009), Rubin and Thomas (2000), Glazerman et al. (2003) Imai and Van Dyk (2004), Abadie and Imbens (2006) and Ho et al. (2007) suggested that adjusting for confounders in an outcome model may significantly improve inference on causal effects.

Thus, model assisted estimation of causal effects is a common practice in causal inference. However, there have been relatively few formal investigations of the consequences of model misspecification for different propensity score methods, and whether results are more sensitive to misspecification of the outcome or treatment assignment model. Previous studies of model misspecification in the context of causal inference have grouped misspecified models in broad ad-hoc categories such as "incorrect model" or "wrong model" (Drake, 1993; Kang and Schafer, 2007; Robins et al., 2007). To our knowledge, this is the first attempt to systematically quantify the degree of model misspecification and evaluate its impact on two of the more commonly used estimation procedures (i.e., propensity score matching and weighting) under different survey designs.

Complex survey designs provide an extra layer of complexity when estimating causal effects. Non-experimental studies often use complex survey data, but there is relatively little guidance on how to incorporate the survey design in propensity score methods. Work by Austin et al. (2016) and Lenis et al. (2017b) extended the use of propensity score matching to complex survey data. Similarly, Ridgeway et al. (2015) provided some insight on how to compute IPTW estimators using complex survey data; however, it is unclear whether model misspecification would have different implications in the complex survey context.

This paper is organized as follows: in Section 4.2, we present key definitions and assumptions needed for the estimation of causal effects in the context of non-

experimental data. Section 4.3 reviews the methods implemented in our simulation study. Section 4.4 contains the details of our simulation study. In Section 4.5, the main results are presented, followed by the discussion and main conclusions in Section 4.6.

4.2 Definitions and Assumptions

4.2.1 The Causal Inference Framework

Traditionally, causal treatment effects are defined using the Rubin Causal Model (RCM) (Rubin, 1974). In the RCM, an individual treatment effect, associated with a binary treatment assignment T, is defined in terms of potential outcomes. For each unit i, $Y_i(t)$ with $t = 0, 1$, represents the outcome that would have been observed if unit i received the treatment t. Thus, the treatment effect for the i^{th} unit is equal to $Y_i(1) - Y_i(0)$. Notice that for any unit i, the pair $(Y_i(0), Y_i(1))$ is not observable - only one of the two potential outcomes is observed. Explicitly, the observed outcome, Y_i, is defined as:

$$Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i) \tag{4.1}$$

Equation (4.1) is referred as the “consistency of the observed outcome assumption" (Hernan and Robins, 2017). Given that the unit level treatment effects cannot be estimated directly, we are often interested in estimating average treatment effects.

At the population level, the most commonly defined average effects are: (1) the population average treatment effect (PATE) and (2) the population average treatment effect on the treated (PATT).

The PATE is defined as average effect across the population:

$$PATE = E\left[Y(1) - Y(0)\right] \tag{4.2}$$

Under randomization of the treatment, units in the treated group and the units in the control group have similar distributions of covariates (observed and unobserved) and potential outcomes. In this way, the average outcome computed among the units in the treated group serves as a good counterfactual for the average outcome computed among the units in the control group. The differences between these two averages is an estimator of the population average treatment effect (PATE).

The PATT is defined as the average causal effect, computed only among those units in the population who were actually treated:

$$PATT = E\left[Y(1) - Y(0)|T = 1\right] \tag{4.3}$$

When the treatment is randomized, it holds that the PATE is equal to the PATT. In non-experimental studies, where the treatment and comparison groups may be quite different from one another on confounders and effects, the PATT and the PATE can be different.

When randomization is not feasible, additional assumptions are required to identify and estimate treatment effects. In particular, a crucial assumption in the estimation of treatment effects is referred to as “ignorability" (Rosenbaum and Rubin, 1983). To further describe the implications of this assumption, we define for all i, $\mathbf{X}_i$ as $q-$dimensional vector of covariates, i.e., $\mathbf{X}_i = (X_{1,i}, \ldots, X_{q,i})$. Ignorability assumes that $\mathbf{X}$ contains all possible confounders: all variables related to treatment assignment and outcome. In other words, given the set of observed covariates $\mathbf{X}$, the treatment assignment is independent of the potential outcomes. The ignorability assumption means that the treatment assignment is random, conditionally on observed characteristics of the units in the sample. This implies that:

$$(Y_i(0); Y_i(1)) \perp\!\!\!\perp T_i | \mathbf{X}_i \tag{4.4}$$

Another key assumption of the RCM is the Stable Unit Treatment Value Assumption (**SUTVA**). The implication of this assumption is twofold: (1) the treatment assignment of any unit does not affect the potential outcomes of other units (often referred to as non-interference) and (2) there is only one version of the treatment, implying that the treatment is comparable across units (Rubin, 1980).

4.2.2 PATT versus SATT

While many researchers are interested in estimating causal effects at a population level, data from a study sample can only be used to truly and consistently estimate a sample ATE (SATE). Estimation of the PATE requires one to have access to data on the full target population of interest, which is rare in practice. The SATE represents the difference in average outcomes if everyone in the survey sample received the treatment versus everyone in the survey sample receiving the control condition. An unbiased estimator of the SATE (SATT) will correctly estimate the PATE (PATT) only when the sample distribution of the relevant variables is similar to its target population counterpart. One sampling design that guarantees this is a simple random sample (SRS), but this kind of sampling technique is hardly ever used. In general, most surveys are the result of complex sampling designs. Therefore, unless survey weights are used to weight the sample back to the population, using sample information to estimate a treatment effect will result in a consistent estimator for the SATE (SATT) but not for the PATE (PATT).

4.3 Propensity Score Methods

In this section we present two commonly used techniques to estimate population causal effects: (1) Propensity Score Matching and (2) Inverse Probability of Treatment Weighting. While we focus on estimating the PATT in this paper, these methods can

also be used to estimate the PATE (Abadie and Imbens, 2006; Ridgeway et al., 2015)

4.3.1 Propensity Score Matching (PSM)

Matching estimators have been widely used in the context of non-experimental studies. They help reduce bias in the estimation of causal effects (Rubin, 1973a), are intuitive and relatively easy to implement. Fundamentally, matching matches individual observations (i.e., comparison to treated units) on a set of observed covariates. The main goal of this matching approach is to generate a new sample (i.e., the matched sample), such that for every treated unit there is (at least) one comparison unit with similar values of observed covariates. The outcome of interest is then compared between the matched treated and matched comparison subjects to estimate the causal effect. One main disadvantage of this procedure resides in the fact that as the number of variables on which units are matched increases, the chances of finding matched pairs with similar observed characteristics decreases exponentially. Thus, matching directly on a set of covariates is only feasible in large samples and/or if a small set of covariates are used in the matching procedure. This is why propensity score matching can be useful. Rosenbaum and Rubin (1983) showed that a similar (or balanced) distribution of the observed characteristics can also be achieved when the matching procedure is based on the propensity score instead of the entire set of observed covariates. Guidelines regarding the implementation of propensity score matching in the context of complex survey data can be found in Austin et al. (2016)

and Lenis et al. (2017b)

4.3.2 Inverse Probability of Treatment Weighting

An alternative approach to estimate causal effects is to compute an inverse probability of treatment weighting (IPTW) estimator. In the context of simple random samples (SRS), an IPTW estimator of the ATT requires, as a first step, the computation of propensity score based weights. The units in the comparison group receive a weight equal to the odds of receiving treatment, while the treated receive a weight equal to one. This serves to weight the comparison group to look similar to the treatment group, thus estimating the ATT (Robins et al., 2000; Harder et al., 2010).

After the propensity score weights are computed, a weighted difference in means (exposed versus unexposed) can be computed in order to estimate the ATT. Furthermore, weighted regression models can be fit to estimate causal effects (Joffe et al., 2004). This approach allows for the estimation of causal effects adjusting for relevant confounders. Ridgeway et al. (2015) developed a strategy to compute an IPTW estimator using complex survey data.

A different weighting strategy needs to be implemented when the goal is to estimate the ATE (Austin, 2011)

4.3.3 Degree of Misspecification (DoM)

Previous work has not explicitly addressed the level of misspecification in the propensity score and outcome model when assessing the impact of misspecification in the estimation of causal effects. In this article we propose a measure of the DoM of a model and explore how the DoM impacts the performance of the estimators considered. Controlling the DoM will allow us to: (1) evaluate how robust the considered estimators are to different levels of DoM and (2) assess whether the same level of DoM in each model (i.e., propensity score and outcome) has the same impact on the performance of the estimators considered.

Throughout this paper, we will use η to represent the DoM for a given model. For a given dependent variable, Z, we define μ_i as the mean of Z conditional on a set of predictors (i.e., $E[Z_i|\mathbf{X}_i]$). We assume that there is a function g^C such that $\mu_i = g^C(\mathbf{X}_i)$. Thus, η can be defined as:

$$\eta = \frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{g}(\mathbf{X}_i) - \widehat{g^C}(\mathbf{X}_i)|}{\sigma_{\widehat{g^C}}} \tag{4.5}$$

Here N represents the population size, $\widehat{g^C}(\mathbf{X}_i)$ is the predicted conditional mean under the correct model specification for unit i with $i = 1, ..., n$, $\widehat{g}(\mathbf{X}_i)$ is the predicted conditional mean under a given model specification for unit i with $i = 1, ..., n$. The

Here $\mathbf{X}_i$ represents the set of predictors. This set can also contain the treatment indicator. Notice that this is implies a slight abuse in notation since in Section 4.2 we defined $\mathbf{X}_i$ as a set of confounders that did not include the treatment indicator

symbol $\sigma_{\widehat{g^C}}$ represents the standard deviation of the predicted conditional means under the correct model specification. Thus when $g = g^C$, we have that $\eta = 0$ and when $g \neq g^C$ $\eta > 0$. Therefore we have that $\eta \in [0, \infty)$, and as η increases, so does the degree of misspecification of a given model.

This measure of DoM has some desirable properties: (1) is **unit independent**, which facilitates the comparisons across different working models and types of dependent variables (e.g., continuous, binary, categorical, etc.), (2) the magnitude is **informative** (i.e., higher values of η are associated higher degree of misspecification), (3) since η is computed in the population, it is not affected by sample size or the survey design.

Notice that η is defined as a parameter in our simulation study and since its computation requires knowledge of the true parametric model, it cannot be used in a real data analysis. Since our simulation study involves the estimation of the propensity score and outcome model, we have a DoM associated with the estimation of the propensity model (η_T) and a DoM related to the outcome model (η_Y).

The DoM of the propensity score model is defined as:

$$\eta_T = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{\pi}_i - \hat{\pi}_i{}^C|}{\sigma_{\hat{\pi}^C}} \tag{4.6}$$

Here $\hat{\pi}_i$ is the predicted probability of being assigned to the treatment group under a given model specification, $\hat{\pi}_i{}^C$ is the predicted probability of being assigned

to the treatment group under the correct model specification and $\sigma_{\hat{\pi}^C}$ is the standard deviation of the predicted probabilities of being assigned to the treatment group under the correct model specification. The DoM associated with the outcome of interest is defined as:

$$\eta_Y = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{Y}_i - \hat{Y}_i^C|}{\sigma_{\hat{Y}^C}} \tag{4.7}$$

Here $\hat{Y}_i$ is the predicted observed outcome under a given model specification, $\hat{Y}_i^C$ is the predicted observed outcome under the correct model specification and $\sigma_{\hat{Y}^C}$ is the standard deviation of the predicted observed outcomes under the correct model specification.

4.3.4 Methods examined

In our simulation study (see Section 4.4) we compute two propensity score based methods to estimate the PATT: (1) propensity score matching and (2) propensity score weighting.

First, w implement a 1-to-1 nearest neighbor matching without replacement. When the sample is obtained using a complex survey design, we follow Lenis et al. (2017b). Since we are assuming a non-response rate of 0% we do not implement the weight transfer described in Lenis et al. (2017b). We do incorporate the survey

weights in the estimation of the propensity score model, since Lenis et al. (2017b) argue that doing so could lead to more efficient estimators of the PATT.

Second, we estimate the PATT by computing an IPTW estimator. When the sample is the result of a complex survey design, we follow Ridgeway et al. (2015). That is, the survey weights are incorporated in the estimation of the propensity score model, and the final weights used in the outcome analysis are constructed by multiplying each survey weight by the propensity score based weights.

When the stratified two stage sample is used to fit the propensity score and outcome models, we use the R package "survey" (Lumley, 2004, 2016) to account for the survey design and weights in the estimation procedure.

4.4 Simulation Study

4.4.1 The Data Generating Mechanism (DGM)

Our simulation study follows closely the one presented by Austin et al. (2016), with some modifications: (1) the PATT and the SATT are different and (2) the degree of misspecification of the working models for the propensity score and outcome can be controlled.

As in Austin et al. (2016), we consider the case of a population of size $N = 1,000,000$, divided into 10 strata. Each strata contains 20 clusters, each composed of $5,000$ units.

We consider **two confounders** X_l with $l = 1, 2$ and a data generating mechanism for the baseline covariates such that: (1) the probability density function is Normal, (2) the covariates are independent (i.e., correlation between the covariates is set equal to 0), (3) the standard deviation for each covariate is equal to 1 and, (4) the means vary across strata and cluster. Explicitly, for each strata j, the mean of the covariates deviates in μ_{lj} from 0, where μ_{lj} are obtained assuming that $\mu_{lj} \sim N\left(0, \tau^{stratum}\right)$. Within each strata, the mean of each cluster (k) deviates from the strata specific mean by μ_{lk}, with $\mu_{lk} \sim N\left(0, \tau^{cluster}\right)$. Thus the distribution of the l^{th} variable, in the j^{th} stratum, among the units of the k^{th} cluster is $X_{ljk} \sim N\left(\mu_{lj} + \mu_{lk},\ 1\right)$. We set $\tau^{stratum} = 0.35$ and $\tau^{cluster} = 0.25$. The values for $\tau^{stratum}$ and $\tau^{cluster}$ are extracted from Austin et al. (2016)

The **treatment assignment** (T_i) model is defined as a Bernoulli random variable $T_i \sim Be\left(p_i\right)$ with $\text{logit}\left(p_i\right) = \alpha_0 + \sum_{l=1}^{2} \alpha_l X_{li} + \delta_d \alpha_3 X_{1i} X_{2i}$ with $\alpha_0 = \log\left(0.20\right)$, $\alpha_1 = \log(2.00)$, $\alpha_2 = \log(2.50)$, and $\alpha_3 = \log(3.00)$. In this model, the multiplier δ_d with $d = 1, .., 11$ allows us to control the degree of misspecification of the working model (see Section 4.4.3) used to estimate the propensity score. The values of δ_d are selected such that the degree of misspecification (DoM) varies from 0.00 to 0.50.

The **potential outcomes** model under **control** is defined as $Y_i\left(0\right) \sim N(\mu_i^0, \sigma^2)$, with $\mu_i^0 = \beta_0 + \sum_{l=1}^{2} \beta_l X_{1i} + \Delta_m(\delta_d)\beta_3 X_{1i} X_{2i} + \sum_{j=2}^{10} \theta_j STR_{ji}$, where $\beta_0 = \log(0.20)$, $\beta_1 = \log(2.50)$, $\beta_1 = -\log(2.00)$, $\beta_3 = \log(4.50)$, $\theta_j = \log(0.50)$ for $j = 2, .., 5$, and $\theta_j = \log(2.00)$ for $j = 6, ..., 10$. The term $\sum_{j=2}^{10} \theta_j STR_{j,i}$ ensures that the PATT and

the SATT will be different. The variable $STR_{j,i}$ is a categorical variable that takes the value 1 if the sample unit i belongs to the j^{th} stratum. The parameter $\Delta(\delta_d)_m$ with $m = 1, ..., 11$ is indexed by δ_d to ensure that for every degree of misspecification in the propensity score model, the degree of misspecification of the outcome model also ranges from 0.00 to 0.50 by 0.05 increments. The potential outcome under **treatment** is defined by $Y_i(1) \sim N(\mu_i^1, \sigma^2)$, with $\mu_i^1 = \mu_i^0 + \gamma$, with $\gamma = \log(3.00)$. Recall that the observed outcome (Y_i) is defined as:

$$Y_i = T_i \times Y_i(1) + (1 - T_i) \times Y_i(0)$$

We model the outcome of interest as a continuous variable for two reasons. First, since the treatment effect is homogeneous, the PATT is equal to γ. Second, having a continuous outcome will allow us fit a model for the outcome of interest in the matched sample that will yield a consistent estimator of the PATT. This is due to the fact, as stated in Austin et al. (2016), "that propensity score methods result in marginal estimates of effect, rather than conditional estimates of effect. When outcomes are continuous, a linear treatment effect is collapsible: the conditional and marginal estimates coincide. When the outcome is binary, regression adjustment in the propensity score matched sample will typically result in an estimate of the odds ratio. The odds ratio (like the hazard ratio) is not collapsible; thus the marginal and conditional estimates will not coincide."

4.4.2 Survey Designs

In our simulation study we consider two sampling schemes: first, we consider a simple random sampling scheme. Under this survey design, $5,000$ units were randomly selected from the population without replacement. Second, we also implement a two stage stratified sample. As mentioned in Section 4.4.1, the target population consists of 10 strata, each with 20 clusters. Within each stratum, 5 clusters are selected randomly without replacement. Within each selected cluster, we draw a random sample without replacement of the final sampling units. Within each stratum, the same number of observations are selected among the sampled clusters. We allocate sample sizes to the 10 strata as follows: 750, 700, 650, 600, 550, 450, 400, 350, 300, and 250. Therefore, the final sample consists of $5,000$ units, which represents 0.5% of the target population. Survey weights are constructed to be equal to the inverse of the selection probability. Strata divide the population in mutually exclusive and exhaustive groups, and clusters within each stratum are randomly selected. Thus every strata is represented in the final sample, but not every cluster. For example, strata could be defined by states, while counties or street blocks define the clusters. In this example, every state will be represented in the final sample but not every county.

For simplicity, we assume a 0% non-response rate (work by Lenis et al. (2017b) explored the consequences of the non-response in the estimation of population causal effects in the context of complex survey data).

We implement $1,000$ iterations in our simulation study. That is, under both sampling schemes $1,000$ samples are drawn from the population.

4.4.3 Analysis models

After the sample is obtained, the following propensity score model is estimated:

$$\text{logit}\left(p_i\right) = a_0 + \sum_{j=1}^{2} a_j X_{ji} \tag{4.8}$$

Notice that by setting $\delta_1 = 0$ (see Section 4.4.1) the working model defined by equation 4.8 will be correctly specified, thus making the degree of misspecification equal to 0 ($\eta_T = 0$). The analysis outcome model is defined by the following equation:

$$m_i = b_0 + \sum_{j=1}^{2} b_j X_{j1} + \sum_{j=2}^{10} b_{j+2} STR_{ji} + b_{13} T_i \tag{4.9}$$

Here m_i represents the model for the mean of the observed outcome, given the confounders, the strata identifier and the treatment assignment. Again, by setting $\Delta_1(\delta_d) = 0$ for all δ_d (see Section 4.4.1), the working model defined by equation 4.9 will be correctly specified, making the associated degree of misspecification equal to 0 ($\eta_Y = 0$). Notice that the working models defined by equations 4.8 and 4.9 include all confounders, thus the assumption of no unmeasured confounders holds. Therefore, the source of the misspecification in both models is the omission of the interaction

term (i.e., X_1X_2).

4.5 Results

In this section we evaluate the performance of the estimator of γ as a function of the degree of misspecification using the following three metrics: (1) percentage bias (in absolute value), (2) empirical coverage of the 95% confidence interval and (3) root mean squared error.

Our main results are summarized in figures 4.1, 4.2 and 4.3. The vertical axis displays the DoM in the outcome model (η_Y) while the horizontal axis shows the DoM in the propensity score model (η_T). The top two panels show the results associated with a SRS while the bottom two panels display the results associated with a two-stage stratified sample. The two panels on the left show the results using the propensity score matching approach and the plots on the right are associated with the IPTW estimator.

Figure 4.1 shows how the percentage of bias (in absolute values) is affected by the DoM in both models. Lighter shades indicate less bias, while darker shades indicate higher levels of bias. Observe from Figure 4.1 that results are similar for the simple random sample (top two panels) and a complex survey design (bottom two panels). As expected, the bias of the estimator increases as the DoM increases in both models. In fact, when the DoM is 0.50 in both models, the bias (in absolute value) can be as

Plots with value labels are available in the appendix.

high as 200%. When the outcome model is correctly specified ($\eta_Y = 0$), both methods yield unbiased estimators of the PATT, regardless of the DoM associated with the propensity score model ($\eta_T \geq 0$). When the propensity score is correctly specified ($\eta_T = 0$) we observe that the IPTW method (right panels) returns an estimator that is unbiased regardless of the level of DoM associated with the outcome model ($\eta_Y \geq 0$). This is due to the fact that the procedure used in the computation of the IPTW estimator yields the doubly robust estimator attributed to Joffe (Robins et al., 2007). Doubly robust estimators (Scharfstein et al., 1999; Kang and Schafer, 2007) yield consistent estimators of the PATT when either the propensity score or the outcome model (but not necessarily both) are correctly specified (i.e., $\eta_T = 0$ or $\eta_Y = 0$). This same result does not hold for the propensity score matching estimator. From Figure 4.1, observe that when the propensity score model is correctly specified (i.e. $\eta_T = 0$) the bias of the matching estimator increases as the DoM of the outcome model also increases (i.e. $\eta_Y \geq 0$). Therefore misspecifying the propensity score model results in smaller biases than misspecifying the outcome model. This result is consistent with the one obtained by Drake (1993).

Figure 4.2 displays the results associated with the empirical coverage of the 95% confidence interval. Lighter shades indicate higher coverage, while darker shades depict lower empirical coverage. Observe that there is a sharp fall in the coverage when the DoM exceeds 0.15 in both models. This is due to the fact that values of DoM higher than 0.15 are associated with bias larger than 10% (see Figure 4.1).

Therefore, this pattern is expected, since the confidence intervals are centered at a value far from the true value of γ.

Figure 4.3 summarizes the results for RMSE. Lighter shades indicate lower values of the RMSE, while darker shades show higher levels of RMSE. Notice that Figure 4.1 and Figure 4.3 display a similar pattern, indicating that there are no significant differences in the efficiency of the estimation procedures.

4.6 Discussion

In this paper, we explore how model misspecification affects the performance of two of the most commonly used methods to estimate the PATT: (1) propensity score matching and (2) IPTW. As noted in Section 4.4.3, an outcome model that adjusts for the confounders was used to estimate the PATT (i.e., γ).

One contribution of this paper is the careful quantitification of model misspecification. In Section 4.3.3 (see equations 4.6 and 4.7) we presented η, a metric of the degree of misspecification for a given model. Given that η is unitless, it can be used to compare the DoM of different models and different types of dependent variables. The fact that η is not affected by the sample size and survey design allowed us to evaluate the performance of the estimators in the context of complex survey data and simple random sampling.

To our knowledge, this is the first attempt to systematically quantify the degree

of model misspecification in the analysis models in order to evaluate its impact in the estimation of causal effects. Still, there are some limitations to our approach. The metric used to quantify the degree of misspecification (i.e., η) is computed at the population level and requires knowing the true model. Future work will focus on providing measures of the DoM that can be computed at a sample level. Additionally, we only explored the consequences of omitting the interaction term. In our simulation study, link functions are correctly specified and all relevant confounders are observed and measured without error. Future work will focus on assessing the impact of other types of model misspecification.

Based on the metric of model misspecification, we evaluated the performance of methods for estimating the PATT in the presence of propensity score and/or outcome model misspecification. Perhaps not surprisingly, but importantly, we found similar results across simple random samples and complex survey sample designs. This is useful guidance for researchers and implies that findings may be similar for a variety of study designs

Both estimation procedures yield similar performance in terms of bias, coverage and RMSE when the outcome model is correctly specified ($\eta_Y = 0$), but the propensity score model is not ($\eta_T \geq 0$). When the propensity score model is correctly specified ($\eta_T = 0$), the IPTW estimator is robust to different degrees of misspecification associated with the outcome model. This is expected given the doubly robust nature of this estimator. When the propensity model is correctly specified ($\eta_T = 0$),

the bias of the propensity score matching estimator increases as the DoM of the outcome model increases ($\eta_Y \geq 0$). Thus, when implementing matching, misspecifying the propensity score model results in smaller biases than misspecifying the outcome model. This confirms the results from Drake (1993). Nevertheless, it is important to keep in mind that true models are rarely known. Thus in practice, it is very likely that both models (i.e., the propensity score and the outcome) will be misspecified (i.e., $\eta_T > 0$ and $\eta_Y > 0$). When this is the case, the performance of both estimation procedures is practically identical, which confirms the results obtained by Kang and Schafer (2007).

In conclusion, as the degree of model specification increases, the performance of the estimators considered worsens. Under the more realistic scenario that both models (i.e., propensity score and outcome) present some degree of misspecification, the performance of the two estimators considered is practically identical. Thus, there is no methodological substitute for well a informed and carefully planned model specification.

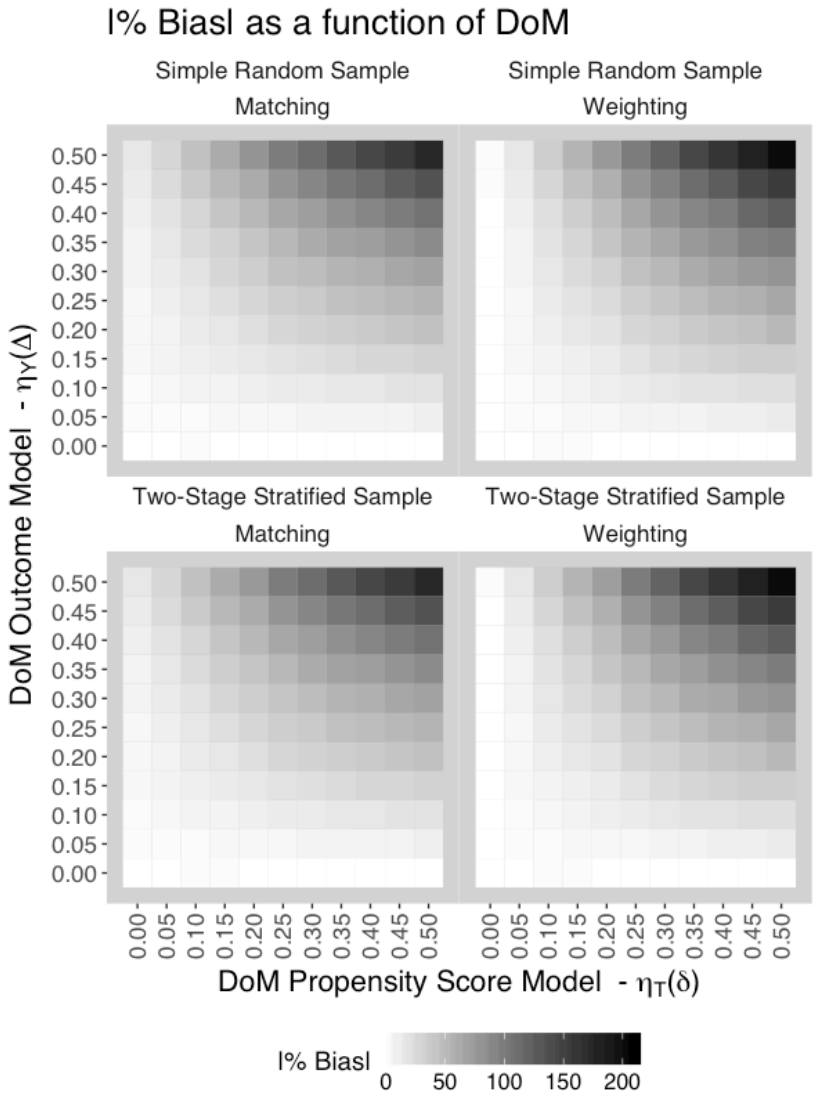

Figure 4.1: |% Bias.| % Bias in absolute value associated with the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_δ), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).

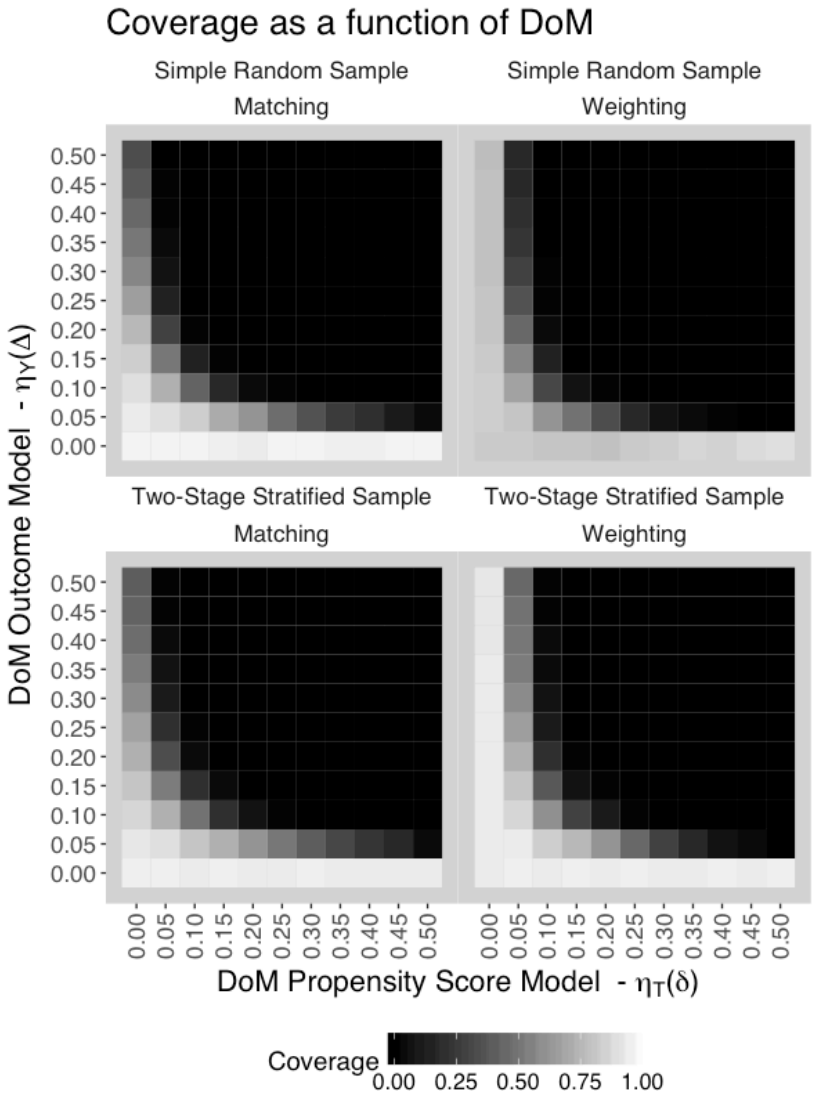

Figure 4.2: |Coverage.| Empirical coverage of the 95 interval in the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_δ), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).

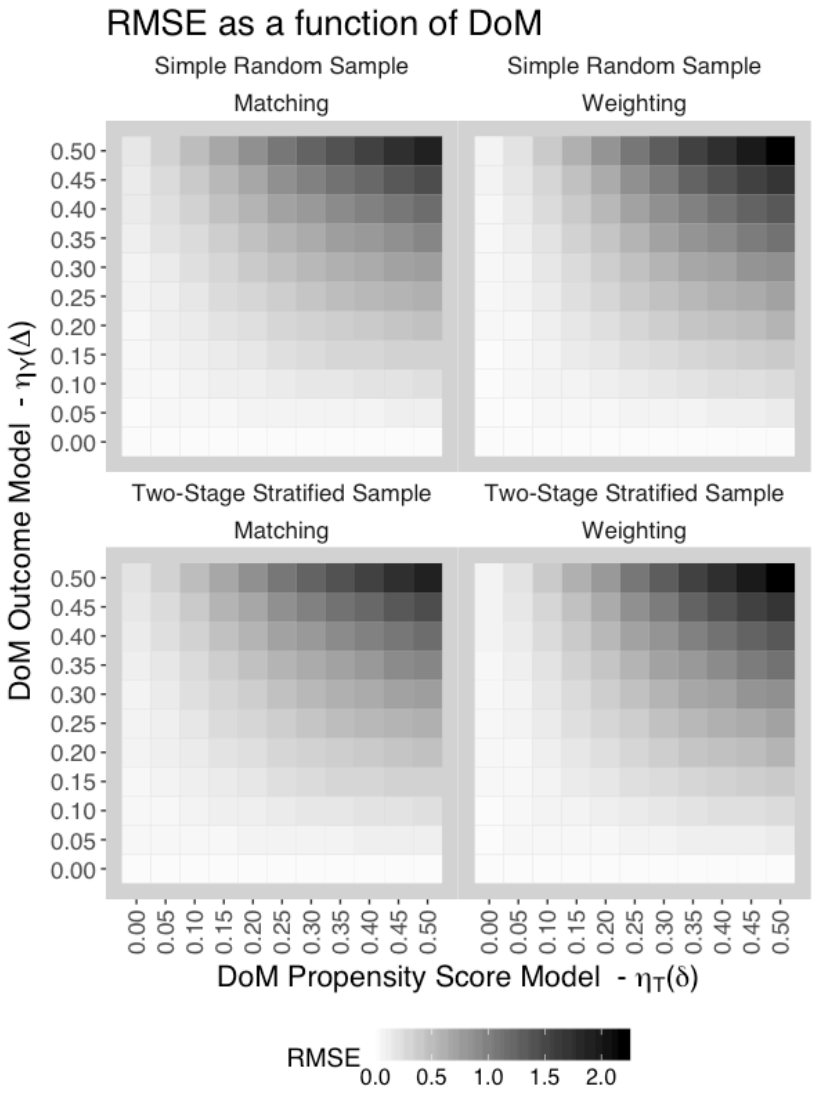

Figure 4.3: **|RMSE.|** RMSE associated with the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_δ), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).

Chapter 5

Discussion

Throughout this article we explored relaxing some of the traditional assumptions used in the estimation of causal effects, in a context of non-experimental data.

First, we relaxed the assumption that all confounders are measured without error by extending the SIMEX methodology to compute a doubly robust estimator of the average treatment effect when a single covariate is measured with error. Furthermore, we presented a more general structure of measurement error which allows for a mean reverting measurement error and has the classical measurement error structure as a special case. This extension is particularly relevant to public health research where measurement error tends to be rule rather than the exception, especially since the use of self-reported data is becoming increasingly common. Variables like, weight, BMI, height and income are some examples of variables, that when are self reported, display a mean-reverting error structure. They main drawback associated with using our more

general structure of measurement error is that the coefficient in the outcome model associated with the covariate measured with error will be inconsistently estimated. Nevertheless in Appendix 6.1 we provided a procedure that solves this problem. There are, nonetheless, limitations associated with this work: (1) the SIMEX extension can only be used when there is a single miss-measured confounder, (2) the estimation procedure relies heavily on parametric assumptions for the outcome and propensity score model and the model chosen for the extrapolation step, and (3) the extension can only handle continuous outcomes. Future work will focus on extending this estimation procedure for different types of outcome variables such as binary and categorical.

In Chapter 3 of this manuscript, we presented a set of guidelines to estimate the average treatment effect among the treated using complex survey data. This work is particularly relevant to public health research since non-experimental data are increasingly being used to estimate causal effects, especially when a randomized trial is infeasible. Large scale, complex survey designs are widely used to estimate causal effects but limited work has been done to create clear and concise guidelines regarding the estimation of causal effects using complex survey data. In this chapter, we provided a formal justification for the weight transfer first proposed by Reardon et al. (2009) and our simulation study, also extended previous work by incorporating a key feature associated with complex survey data: non-response. Up to this point, the work that explored the implementation of causal inference methods in the context of complex survey data (Ridgeway et al., 2015; Austin et al., 2016) failed to incorporate

this feature in their simulation studies. We considered different non-response models and concluded that when the non-response depends on the exposure, implementing the weight transfer yielded a less biased estimator of the population average treatment effect on the treated. We have also showed that when estimating causal effects using complex survey data, population balance (i.e., standardized mean difference) should be computed. That is, the survey weights should be included in the computation of measures of balance. It is important to notice that our conclusions hold for matching without replacement. Future work will extend this procedure to different propensity score matching algorithms and will also focus on generalizing different measures of balance such that survey weights can be incorporated in their computation.

Finally, in Chapter 4, we explored the consequences of model misspecification in the estimation of causal effects. We introduced a metric to quantify the degree of misspecification in a given model. To our knowledge this is the first effort to systematically quantify the degree of model misspecification. Not surprisingly, we found similar results across simple random samples and samples with a complex survey design. When the outcome model was correctly specified, we noticed that both estimation procedures yielded similar performance in terms of bias, coverage and RMSE for different DoM in the propensity score model. Nevertheless, when the propensity score model was correctly specified, we observed that the IPTW estimator was robust to different degrees of misspecification associated with the outcome model. This was expected given the doubly robust nature of this estimator. When consider-

ing the propensity score matching estimator, we observed that when the propensity model was correctly specified, its bias increased as the DoM of the outcome model increased. Thus, when using propensity score matching, misspecifying the propensity score model resulted in smaller biases than misspecifying the outcome model. This confirms the results reached by Drake (1993). Under the more realistic scenario that both models (i.e., propensity score and outcome) present some degree of misspecification the performance of the two estimators considered was practically identical. Thus, we found that there is no methodological substitute for an informed and carefully planned model specification.

Throughout this manuscript we relaxed some of the assumptions associated with the estimation of causal effects, in a context of non-experimental data. Furthermore, we explored the consequences of model misspecification in the performance of two widely used estimators of the ATT. We hope that this work will help inform the discussion on how to estimate causal effects, as public health research relies more frequently on non-experimental data.

Chapter 6

Appendix

6.1 A Motivating Example (Appendix A, Chapter 2).

Consider the simple case of a linear regression model with only two covariates X and Z and an outcome Y. In this motivating example, X is measured with error. The measurement error structure is the same as the one presented in Section 2.1 of the main document. To assess the consequences of having a variable measured with error we express X_i as a function of W_i. Explicitly

$$X_i = \frac{W_i - \sigma_\epsilon \epsilon_i + \tau_1 E(X_i)}{1 + \tau_1} \tag{6.1}$$

Replacing (6.1) in the outcome model, we obtain that:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n-1} \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & \frac{W_1-\sigma_\epsilon\epsilon_i+\tau_1 E(X)}{1+\tau_1} & Z_1 \\ \vdots & \vdots & \vdots \\ 1 & \frac{W_{n-1}-\sigma_\epsilon\epsilon_i+\tau_1 E(X)}{1+\tau_1} & Z_{n-1} \\ 1 & \frac{W_n-\sigma_\epsilon\epsilon_i+\tau_1 E(X)}{1+\tau_1} & Z_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_X \\ \beta_Z \end{bmatrix} + \begin{bmatrix} \upsilon_1 \\ \vdots \\ \upsilon_{n-1} \\ \upsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n-1} \\ Y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & W_1 & Z_1 \\ \vdots & \vdots & \vdots \\ 1 & W_{n-1} & Z_{n-1} \\ 1 & W_n & Z_n \end{bmatrix}}_{W'} \underbrace{\begin{bmatrix} \beta_0 + \beta_X \frac{\tau_1 E(X_1)}{1+\tau_1} \\ \frac{\beta_X}{1+\tau_1} \\ \beta_Z \end{bmatrix}}_{\beta'} + \underbrace{\begin{bmatrix} \upsilon_1 - \beta_X \frac{\sigma_\epsilon \epsilon_1}{1+\tau_1} \\ \vdots \\ \upsilon_{n-1} - \beta_X \frac{\sigma_\epsilon \epsilon_{n-1}}{1+\tau_1} \\ \upsilon_n - \beta_X \frac{\sigma_\epsilon \epsilon_n}{1+\tau_1} \end{bmatrix}}_{\mu}$$

Then the linear regression estimator for β' is computed as:

$$\begin{aligned} \widehat{\beta'} &= \left[W'^T W\right]'^{-1} W'^T Y \\ &= \beta' + \left[\frac{1}{n} W'^T W'\right]^{-1} \frac{1}{n} W'^T \mu \end{aligned}$$

Notice that since $\frac{1}{n} W^T \mu$ does not converge to a null vector, ignoring the measurement error will lead to inconsistent and biased estimators of the regression coefficients. Furthermore, SIMEX will provide a consistent estimator of β' but not β. Thus when the model for the conditional mean of the outcome is linear; the treatment indicator is measured without error and the model includes a covariate that is mismeasured,

the treatment effect will be consistently estimated after applying SIMEX but the coefficient associated with the surrogate will not. Nonetheless there are scenarios where the coefficient associated with the mismeasured variable may be of interest. In this section we propose a simple procedure to obtain a consistent estimator of the coefficient associated with X (the covariate measured with error) and valid confidence intervals. By Carroll et al. (1996) we know that

$$\sqrt{n}\left(\frac{\widehat{\beta_X}}{1+\tau_1}-\frac{\beta_X}{1+\tau_1}\right)\xrightarrow{D}N\left(0,\Lambda_1\right)$$

Recall that in the validation sample, both X and W are observed. Let $j=1,...,m$ index the units in our independent external validation sample. Let W_j represent, for unit j, the measurement of X_j. It is important to note, in this validation sample both W_j and X_j are observed. Since we can rewrite W_i as $W_i=-\tau_1 E\left(X_i\right)+(1+\tau_1)X_i+\sigma\epsilon_i$, thus $(1+\tau_1)$ can be estimated using a simple linear regression model. Thus we can have that $\sqrt{m}\left(\widehat{1+\tau_1}-1-\tau_1\right)\xrightarrow{D}N\left(0,\Lambda_2\right)$ and under the assumption that $\frac{m}{n}\to k$, we can conclude that $\sqrt{n}\left(\widehat{1+\tau_1}-1-\tau_1\right)\xrightarrow{D}N\left(0,\frac{1}{k}\Lambda_2\right)$. Since the main and the validation samples are independent it holds that

$$\sqrt{n}\left[\begin{pmatrix}\widehat{\frac{\beta_X}{1+\tau_1}}\\ \widehat{1+\tau_1}\end{pmatrix}-\begin{pmatrix}\frac{\beta_X}{1+\tau_1}\\ 1+\tau_1\end{pmatrix}\right]\xrightarrow{D}N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\Lambda_1 & 0\\ 0 & \frac{1}{k}\Lambda_2\end{pmatrix}\right]$$

Thus we obtain that $\sqrt{n}\left(\widehat{\beta_X}-\beta_X\right)\xrightarrow{D}N\left(0,\Lambda\right)$, with $\Lambda=\left(\frac{\beta_X}{1+\tau_1}\right)^2\frac{1}{k}\Lambda_2+(1+$

$\tau_1)^2\Lambda_1$. A similar procedure can be implemented to obtain a consistent estimator of the intercept in the outcome model.

6.2 Simulation Set-Up (Appendix B, Chapter 2).

To evaluate the performance of our estimator we conduct a simulation study to compare bias, mean squared error (MSE) and coverage of three different estimators of the treatment effect Δ: (1) the estimator obtained by using X, the true measure of the covariate, (2) a naive estimator, which ignores the measurement error and simply uses W, and (3) the SIMEX estimator for the treatment effect. The three methods implement a doubly robust approach using propensity score weights. A total of 1000 simulation iteration were implemented for each set of simulation parameters. For each simulation we generated a main sample of $n = 2500$ units and independent validation sample of $m = 500$ units. Within each simulation the SIMEX estimator was computed using $B = 100$ and $\Lambda = \{\lambda = 0.02 + l \times 0.04 : l = 0, ..., 49\}$. Finally, we set $\mathcal{G}(\vartheta, \lambda)$ as a quadratic function; explicitly $E\left(\hat{\Theta}_\lambda|\lambda\right) = \omega_0 + \omega_1\lambda + \omega_2\lambda^2$.

6.2.1 The data generating process

6.2.1.1 The Covariates

We assume that there are only two relevant covariates in the propensity score and in the conditional mean model, namely X and Z, where $X \in \mathbb{R}$ and $Z \in \mathbb{R}$. We generate X_i and Z_i from a bivariate normal distribution with correlation ρ, namely:

$$\begin{pmatrix} X_i \\ Z_i \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

We consider three values for ρ: 0.3, 0.5 and 0.7.

6.2.1.2 The Treatment Assignment

Treatment is assigned according to the following rule:

$$T_i | X_i, Z_i \sim Bernoulli \left(\frac{e^{\alpha_0 + \alpha_1 X_i + \alpha_2 Z_i}}{1 + e^{\alpha_0 + \alpha_1 X_i + \alpha_2 Z_i}} \right)$$

Where α_0 is fixed at 0.5, α_Z is fixed at 0.6, and α_X takes two values: 0.5 and 1.0.

6.2.1.3 The Outcome

The data generating process for the potential outcomes is $Y_i(T_i) = \beta + \Delta T_i + \beta_X X_i + \beta_Z Z_i + \zeta_i$ where we generate independently identically distributed errors ζ_i from a standard normal distribution. Again the parameter of interest is Δ, which is

set equal to 1. We set β equal to 0.5, β_X equal to 0.5 and β_Z equal to 0.6. The observed outcome Y is constructed from the simulated data as $Y_i = T_i \times Y_i(T_i) + (1 - T_i) \times Y_i(T_i)$

6.2.1.4 Measurement Error

Throughout this article we assume a non-traditional classical measurement error structure defined in Section 2.1 of the main document. We set the value of σ^2 to 0.35 and we consider five possible values for $\tau_1 : -0.25, -0.20, -0.15, -0.10$ and 0.00. Table 6.1 summarizes all the parameters used in the simulation study.

Table 6.1: Parameters used in the simulation study.

	Parameter	Value		Parameter	Value
Sample	N_{sim}	1000	W	$\tau_1^{(1)}$	0.25
	m	500		$\tau_1^{(2)}$	0.20
	n	2500		$\tau_1^{(3)}$	0.15
T	α	0.5		$\tau_1^{(4)}$	0.10
	α_Z	0.6		$\tau_1^{(5)}$	0.05
	α_X (small)	0.5			
	α_X (large)	1.0			
(X, Z)	low ρ	0.3		σ	0.34
	medium ρ	0.5	Y	β	0.5
	high ρ	0.9		β_Z	0.6
SIMEX	B	100		β_X	0.5
	Λ	$\{\lambda = 0.02, 0.06, ..., 1.98\}$		ξ	1

N_{sim} represents the number of iterations.

6.3 Non-response mechanisms (Appendix A, Chapter 3).

Traditionally, missing data mechanisms are grouped in three categories: (1) Missing Completely at Random (MCAR), (2) Missing and Random (MAR) and (3) Missing not at Random (MNAR). Under a MCAR mechanism, the probability that one observation will have missing information, is completely random. In other words, there is no relationship between the propensity of the data to be missing and the values of the variables in the data set. When the non-response follows a MAR mechanism, the propensity of the data to be missing is random, conditional on the set of observed variables. In other words the observed values of the available data, can predict the probability of one observation to have missing information. Finally when the non-response is MNAR, the probability of having missing information depends on unobserved variables. That is, even after accounting for the observed variables available in the data, the propensity of the data to be missing is not random.

6.4 Non-response and Survey Weights (Appendix B, Chapter 3).

The survey weights, ω, are equal to the inverse of the probability of being observed in the sample, formally:

$$\omega = \frac{1}{p} = \frac{1}{f_{SR|(\mathbf{X},T)}\left(SR = 1|\mathbf{X}, T\right)} \tag{6.2}$$

Notice that this definition allows for a non-response rate different from 0 and different non-response mechanisms. To see this, consider the case where S and R are independent conditional on $(\mathbf{X}, T)$. Then it holds that $f_{SR|(\mathbf{X},T)}\left(SR = 1|\mathbf{X}, T\right)$ it's equal to $f_{S|(\mathbf{X},T)}\left(S = 1|\mathbf{X}, T\right)$ times $f_{R|(\mathbf{X},T)}\left(R = 1|\mathbf{X}, T\right)$, this last term models the non-response mechanism. Notice that if $f_{R|(\mathbf{X},T)}\left(R = 1|\mathbf{X}, T\right) = 1$ for all $(\mathbf{x}, t)$ in $(\mathbf{X}, T)$ then the non-response rate is 0. If $f_{R|(\mathbf{X},T)}\left(R = 1|\mathbf{X}, T\right) = f_R\left(R = 1\right)$ the non-respond mechanism is MCAR. Finally, the non response could be MAR and NMAR depending on whether the all the elements in $(\mathbf{X}, T)$ are observed. If every element in $(\mathbf{X}, T)$ is available to estimate the probability of non-response, then the non-response mechanism is MAR, otherwise the non-response process is NMAR. In this way, ω (the final observed sampling weight) is a combination of the survey weights associated with the sampling design itself but also incorporates corrections associated with non-response.

6.5 Estimating the PATT (Appendix C, Chapter 3.

Here we incorporate the weight transfer to estimate the PATT. The PATT will be estimated as the difference of the weighted mean of the observed outcomes of the treated and their matched comparison units. This estimator of the PATT makes the use of the weights explicit, nevertheless it is important to recall that a outcome model can be defined and the weights can be incorporated in its estimation. Under the assumption that a $k:1$ matching procedure was implemented it holds that for every treated unit j with $j = 1, 2, \ldots, n_T = \sum_{i=1}^{N} SR_i \times T_i$, we have $h(j) = 1, 2, \ldots k$ comparison units. Thus the $PATT$ can be computed by

$$
\begin{aligned}
\widehat{PATT} &= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} \omega_j^t(\mathbf{x})} \\
&= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{\sum_{h(1)=1}^{k} \omega_1^t(\mathbf{x}) + \sum_{h(2)=1}^{k} \omega_2^t(\mathbf{x}) + ... + \sum_{h(n_T)=1}^{k} \omega_{n_T}^t(\mathbf{x})} \\
&= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{k\omega_1^t(\mathbf{x}) + k\omega_2^t(\mathbf{x}) + ... + k\omega_{n_T}^t(\mathbf{x})} \\
&= \frac{\sum_{j=1}^{n_T} y_j \times \omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} - \frac{\sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times \omega_j^t(\mathbf{x})}{k \sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \\
&= \sum_{j=1}^{n_T} \left[y_j \times \frac{\omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \right] - \sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} \left[y_{h(j)} \times \frac{\omega_j^t(\mathbf{x})}{k \sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \right] \\
&= \sum_{j=1}^{n_T} y_j \times W_j^t - \sum_{j=1}^{n_T} \sum_{h(j)=1}^{k} y_{h(j)} \times W_j^c
\end{aligned}
$$

Defining

$$
\begin{aligned}
W_j^t &= \frac{\omega_j^t(\mathbf{x})}{\sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})} \\
W_j^c &= \frac{\omega_j^t(\mathbf{x})}{k \sum_{j=1}^{n_T} \omega_j^t(\mathbf{x})}
\end{aligned}
$$

We can conclude that

$$
W_j^c = \frac{1}{k} W_j^t
$$

Notice that each of the n_T treated units receives a weight of W_j^t and each of the comparison units a weight of $\frac{1}{k}W_j^t$. Interestingly, when this weight transfer is implemented and then the PATT is estimated, we find that the estimation procedure assigns to each unit of the comparison group a final weight that is proportional to the final weight received by the treated unit to which they have been matched to. Such proportion is defined by the number of comparison units used to matched each treated unit. Notice that a similar result is obtained when considering simple random samples, see Stuart (2010).

Simulation Study (Appendix D, Chapter 3).

As in Austin et al. (2016), we consider the case of a population of size $N = 1,000,000$. There are 10 strata in the population, each stratum has a total of $100,000$ observations. Within each strata there are 20 clusters, each of is composed of $5,000$ units. There are **six covariates** X_l with $l = 1, ..., 6$ and the data generating mechanism for the baseline covariates is such that: (1) the probability density function is normal, (2) the covariates are independent (i.e., correlation between any pair of covariates is set equal to 0), (3) the standard deviation, across all the covariates, is equal to 1 and (4) the means vary across strata and cluster. More explicitly, for each strata (j), the mean of the covariates deviates in μ_{lj} from 0, where μ_{lj} are obtained assuming that $\mu_{lj} \sim N(0, \tau^{stratum})$. Within each strata, the mean of each cluster (k)

deviates from the strata specific mean by μ_{lk}, with $\mu_{lk} \sim N\left(0, \tau^{cluster}\right)$. Thus the distribution of the l^{th} variable, in the j^{th} stratum, among the units of the k^{th} cluster is $X_{l,ijk} \sim N\left(\mu_{lj} + \mu_{lk},\ 1\right)$. We set $\tau^{stratum} = 0.35$ and $\tau^{cluster} = 0.25,\ 0.15,\ 0.05$. Each value of $\tau^{cluster}$ defines a different scenario. Unless otherwise specified, values of the population coefficients are the ones used by Austin et al. (2016)

The **treatment assignment** (T_i) model is defined as a Bernoulli random variable $T_i \sim Be\left(p_i\right)$ with $logit\left(p_i\right) = \alpha_0 + \sum_{l=1}^{6} \alpha_l X_{l,i}$ with $\alpha_0 = log\left(\frac{0.3290}{0.9671}\right)$, $\alpha_1 = log(1.10)$, $\alpha_2 = log(1.25)$, $\alpha_3 = log(1.50)$, $\alpha_4 = log(1.75)$, $\alpha_5 = log(2.00)$ and $\alpha_6 = log(2.50)$.

The **potential outcomes** models are defined as $Y_i\left(0\right) = \beta_0 + \sum_{l=1}^{6} \beta_l X_{1,i} + \epsilon$ with $\epsilon \sim N(0,1)$ and $\beta_0 = 0$, $\beta_1 = 2.50$, $\beta_1 = -2.00$, $\beta_3 = 1.75$, $\beta_4 = -1.25$, and $\beta_6 = 1.10$. The potential outcome under treatment is defined by $Y_i\left(1\right) = Y_i\left(0\right) + \delta_0 + \delta_1 \sum_{l=1}^{3} \beta_l X_{l,i} + \sum_{j=1}^{10} \eta_j STR_{j,i}$ with $\delta_0 = 1$ and $\delta_1 = 0.2$. The term $\sum_{j=1}^{10} \eta_j STR_{j,i}$ is the first departure form the simulation set-up design by Austin et al. (2016). This additional term allows us to control how different the PATT and the SATT are. The variable $STR_{j,i}$ is a categorical variable that takes the value 1 if the sample unit i belongs to the j^{th} stratum. For each of the three scenarios we consider six different values for the vector of parameters $(\eta_1, ..., \eta_{10})$ such that $\left(\frac{SATT}{PATT} - 1\right) \times 100$ takes roughly the values $-50\%, -40\%, -30\%, -20\%, -10\%$ and 0% . In addition to a continuous outcome Austin et al. (2016) also considered a dichotomous outcome; in our article, we restrict our attention to continuous outcomes. We define an indicator variable R_m with $m = 1, 2, 3, 4$ which takes the value 1 is the unit responded and

0 otherwise. We consider the following non-response cases: No-missing data (**NM**), $R_{1i} = 1$ for all i. Missing at Random (**MAR**): the non-response rate depends on the six baseline covariates, explicitly we assume that $R_{3i} \sim Be(p_{3i})$ with $logit\,(p_{3i}) = \gamma_0 + \sum_{l=1}^{6} \gamma_l X_{l,i}$ and $\gamma_0 = -log(0.030)$, $\gamma_1 = -log(1.10)$, $\gamma_2 = -log(1.25)$, $\gamma_3 = -log(1.50)$, $\gamma_4 = -log(1.75)$, $\gamma_5 = -log(2.00)$, $\gamma_6 = -log(2.50)$. Missing at Random with an additional covariate X_7 (**MARX**): the non-response rate depends on the baseline covariates but additionally, depends on a covariate X_7 that is not observed in the final sample, but affect the response rate. Formally,$R_{3i} \sim Be(p_{3i})$ with $logit\,(p_{3i}) = \gamma_0 + \sum_{l=1}^{7} \gamma_l X_{l,i}$ where $\gamma_7 = -log(2.50)$. This non-response mechanism, aims to model the situation in which the survey weights can are constructed using information that is only available to the survey team (i.e., X_7), but not available to the final user (e.g., number of contact attempts). The data generating mechanism for the covariate X_7 is the same as the one for the baseline covariates. The final non-response mechanism consider is Missing at Random where the non-response depends on the baseline covariates and the treatment assignment (**MART**). Explicitly $R_{4i} \sim Be(p_{4i})$ with $logit\,(p_{3i}) = \gamma_0 + \sum_{l=1}^{6} \gamma_l X_{l,i} + \Delta T_i$ and $\Delta = -2$.

6.6 Plots with data labels (Appendix A, Chapter 4).

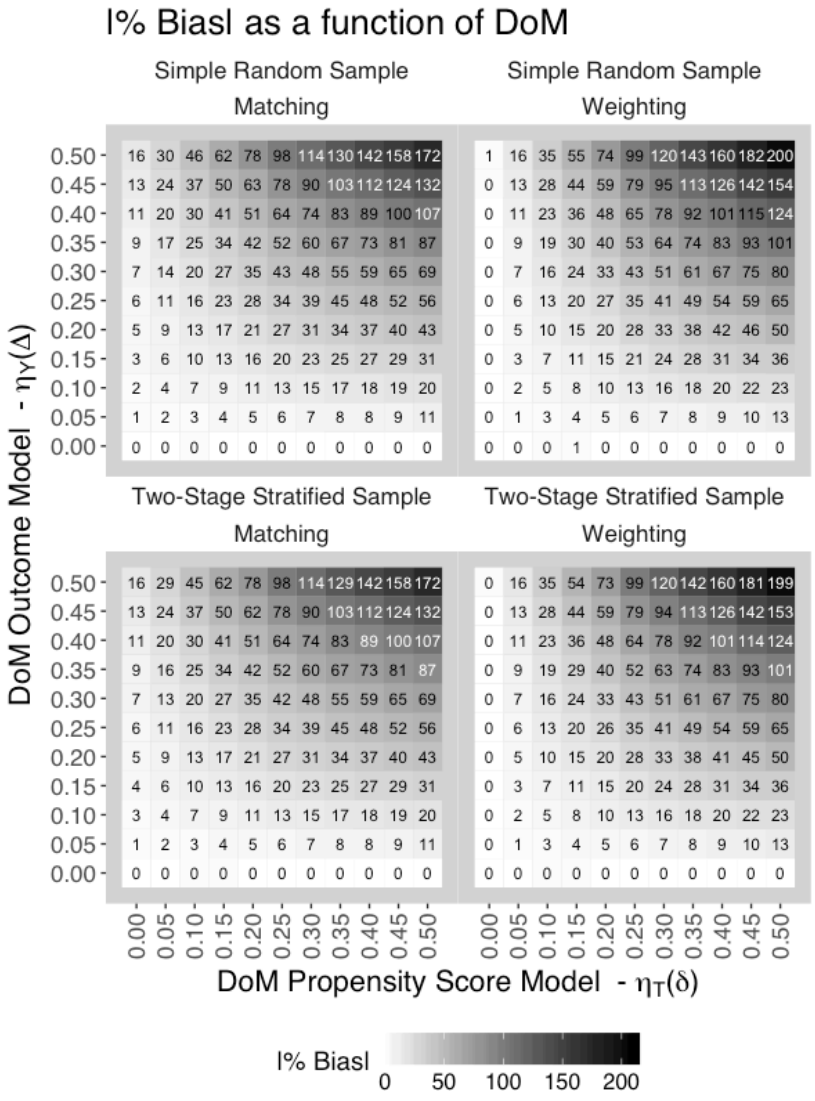

Figure 6.1: |% Bias.| % Bias in absolute value associated with the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_δ), and (2) the Outcome Model. ($\eta_{\Delta(\delta)}$) (simulation study).

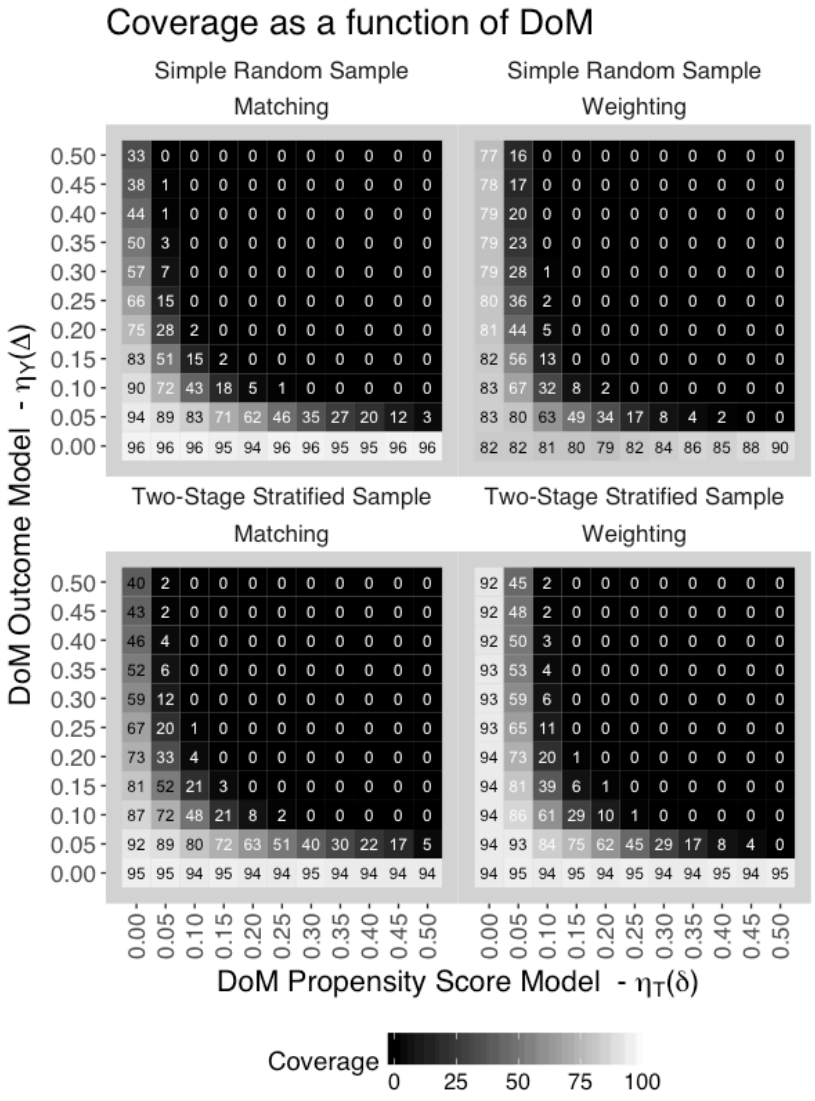

Figure 6.2: |Coverage.| Empirical coverage of the 95 interval in the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_δ), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).

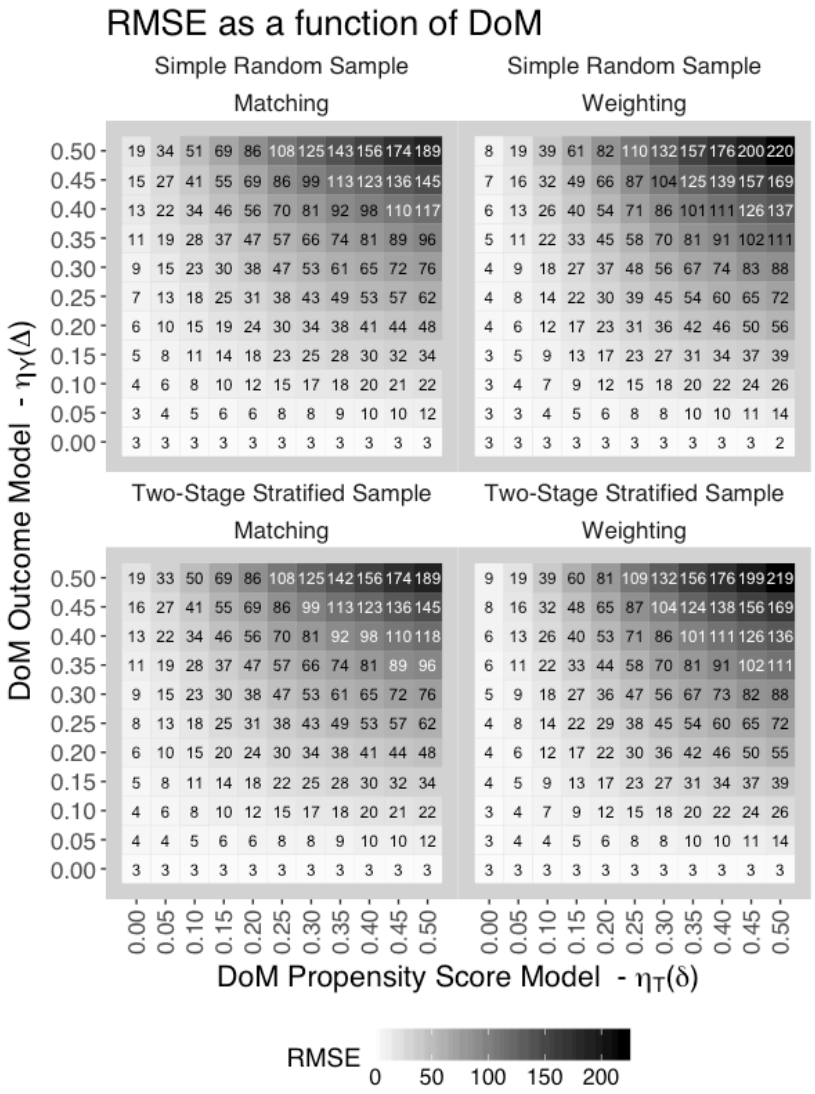

Figure 6.3: **|RMSE.|** RMSE associated with the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_δ), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).

Bibliography

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Akee, R. (2011). Errors in self-reported earnings: The role of previous earnings volatility and individual characteristics. *Journal of Development Economics*, 96(2):409–421.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

BIBLIOGRAPHY

Austin, P. C., Jembere, N., and Chiu, M. (2016). Propensity score matching and complex surveys. *Statistical Methods in Medical Research*, page 0962280216658920.

Bound, J. and Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*, pages 1–24.

Brunell, T. L. and DiNardo, J. (2004). A propensity score reweighting approach to estimating the partisan effects of full turnout in american presidential elections. *Political Analysis*, 12(1):28–45.

Carpenter, R. (1977). Matching when covariables are normally distributed. *Biometrika*, pages 299–307.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the simex estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433):242–250.

Cochran, W. G. (1977). Sampling techniques. 1977. *New York: John Wiley and Sons*.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446.

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International journal of epidemiology*, 35(4):1074–1081.

BIBLIOGRAPHY

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, 89(428):1314–1328.

Dahl, D. B. (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2.

Dowle, M., Srinivasan, A., Short, T., with contributions from R Saporta, S. L., and Antonyan, E. (2015). *data.table: Extension of Data.frame*. R package version 1.9.6.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, pages 1231–1236.

Edwards, J. K., Cole, S. R., Westreich, D., Crane, H., Eron, J. J., Mathews, W. C., Moore, R., Boswell, S. L., Lesko, C. R., Mugavero, M. J., et al. (2015). Multiple imputation to account for measurement error in marginal structural models. *Epidemiology*, 26(5):645–652.

Frölich, M. (2007). Propensity score matching without conditional independence assumption—with an application to the gender wage gap in the united kingdom. *The Econometrics Journal*, 10(2):359–407.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, pages 153–164.

BIBLIOGRAPHY

Glazerman, S., Levy, D. M., and Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, 589(1):63–93.

Goetghebeur, E. and Vansteelandt, S. (2005). Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research*, 14(4):397–415.

Goodman, E. and Whitaker, R. C. (2002). A prospective study of the role of depression in the development and persistence of adolescent obesity. *Pediatrics*, 110(3):497–504.

Grace, Y. Y. (2008). A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*, 9(3):501–512.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey methodology*, volume 561. John Wiley & Sons.

Guo, Y., Little, R. J., and McConnell, D. S. (2012). On using summary statistics from an external calibration sample to correct for covariate measurement error. *Epidemiology*, 23(1):165–174.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

BIBLIOGRAPHY

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793.

Harder, V. S., Stuart, E. A., and Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3):234.

Harris, K. M., Halpern, C., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and J.R., U. (2009). The national longitudinal study of adolescent to adult health: Resear design. *Availabe at: http://www.cpc.unc.edu/projects/addhealth/design Accesed May 15, 2015*.

Heckman, J. J. and Todd, P. E. (2009). A note on adapting propensity score matching and selection models to choice based samples. *The econometrics journal*, 12(s1):S230–S234.

Heid, I., Küchenhoff, H., Miles, J., Kreienbrock, L., and Wichmann, H. (2004). Two dimensions of measurement error: classical and berkson error in residential radon exposure assessment. *Journal of Exposure Science and Environmental Epidemiology*, 14(5):365–377.

Hernan, M. A. and Robins, J. M. (2017). *Causal Inference*. Boca Raton: Chapman & Hall/CRC. Forthcoming.

BIBLIOGRAPHY

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4):272–279.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539.

BIBLIOGRAPHY

Keller, B. and Tipton, E. (2016). Propensity score analysis in ra software review. *Journal of Educational and Behavioral Statistics*, page 1076998616631744.

Korn, E. L. and Graubard, B. I. (1995a). Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 263–295.

Korn, E. L. and Graubard, B. I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49(3):291–295.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174.

Lenis, D., Ebnesajjad, C. F., and Stuart, E. A. (2017a). A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error. *Biostatistics*, 18(2):325–337.

Lenis, D., Nguyen, T. Q., Dong, N., and Stuart, E. A. (2017b). It's all about balance: Propensity score matching in the context of complex survey data. *Under Review*.

Little, R. (2003). The bayesian approach to sample survey inference. *Analysis of Survey Data*, pages 49–57.

Little, R. J. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326.

BIBLIOGRAPHY

Lockwood, J. and McCaffrey, D. (2015). Simulation-extrapolation for estimating means and causal effects with mismeasured covariates. *Observational Studies*, 1:241–290.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19. R package verson 2.2.

Lumley, T. (2016). survey: analysis of complex survey samples. R package version 3.32.

McCaffrey, D. F., Lockwood, J., and Setodji, C. M. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, page ast022.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

Morgan, P. L., Frisco, M. L., Farkas, G., and Hibel, J. (2008). A propensity score matching analysis of the effects of special education services. *The Journal of special education*.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Stat Sci*, 5:463–472.

of Education. Institute of Education Sciences. National Center for Education Statistics., U. S. D. (2011). Early childhood longitudinal study [united states]: Kindergarten class of 1998-1999, kindergarten-eighth grade full sample. icpsr28023-v1. ann arbor, mi: Inter-university consortium for political and social research [distributor]. http://doi.org/10.3886/ICPSR28023.v1http://doi.org/10.3886/ICPSR28023.v1.

Pettersen, B. J., Anousheh, R., Fan, J., Jaceldo-Siegl, K., and Fraser, G. E. (2012). Vegetarian diets and blood pressure among white subjects: results from the adventist health study-2 (ahs-2). *Public health nutrition*, 15(10):1909–1916.

Plankey, M. W., Stevens, J., Fiegal, K. M., and Rust, P. F. (1997). Prediction equations do not eliminate systematic error in self-reported body mass index. *Obesity research*, 5(4):308–314.

Reardon, S. F., Cheadle, J. E., and Robinson, J. P. (2009). The effect of catholic schooling on math and reading development in kindergarten through fifth grade. *Journal of Research on Educational Effectiveness*, 2(1):45–87.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., and Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, 3(2):237–249.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Perfor-

mance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559.

Robins, J. M. (2003). General methodological considerations. *Journal of Econometrics*, 112(1):89–106.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rosner, B., Spiegelman, D., and Willett, W. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case

of multiple covariates measured with error. *American Journal of Epidemiology*, 132(4):734–745.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

BIBLIOGRAPHY

Rubin, D. B. and Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics*, pages 1814–1826.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585.

Saint-Maurice, P. F., Welk, G. J., Beyler, N. K., Bartee, R. T., and Heelan, K. A. (2014). Calibration of self-report tools for physical activity research: the physical activity questionnaire (paq). *BMC public health*, 14(1):1.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Steiner, P. M., Cook, T. D., and Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2):213–236.

Stommel, M. and Schoenborn, C. A. (2009). Accuracy and usefulness of bmi measures

based on self-reported weight and height: findings from the nhanes & nhis 2001-2006. *BMC public health*, 9(1):1.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology*, 162(3):279–289.

Tillé, Y. and Matei, A. (2015). *sampling: Survey Sampling*. R package version 2.7.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., and Najarian, M. (2009). Early childhood longitudinal study, kindergarten class of 1998-99 (ecls-k): Combined user's manual for the ecls-k eighth-grade and k-8 full sample data files and electronic codebooks. nces 2009-004. *National Center for Education Statistics*.

Webb-Vargas, Y., Rudolph, K. E., Lenis, D., Murakami, P., and Stuart, E. A. (2015). Applying multiple imputation for external calibration to propensty score analysis.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wingood, G. M., DiClemente, R. J., Harrington, K., and Davies, S. L. (2002). Body

image and african american females' sexual health. *Journal of women's health & gender-based medicine*, 11(5):433–439.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of data Science*, 4(1):67–91.